The use of alternative preference elicitation methods in complex discrete choice experiments

Hong Il Yoo[a,*], Denise Doiron[b]

[a] Durham University Business School, Durham University, United Kingdom. h.i.yoo@durham.ac.uk

[b] School of Economics, University of New South Wales, Australia. d.doiron@unsw.edu.au

Acknowledgments:
This work was supported by Discovery Project Grant DP0881205 from the Australian Research Council. We thank the project research team: Jane Hall, Debbie Street and Patsy Kenny. We are especially thankful to Agne Suziedelyte for outstanding research assistance. We also wish to thank participants at the AHES Conference 2011, the Econometric Society Australasian Meeting 2012, the Australian Conference of Economists 2012, the Economics of the Health Workforce Conference 2013, and workshops at UNSW. We thank Editor Richard Frank and two anonymous referees for helpful and constructive comments.

*Corresponding author at: Durham University Business School, Mill Hill Lane, Durham DH1 3LB, United Kingdom. Ph: +44 191 334 5544. Fax: +44 191 334 5201. Email: h.i.yoo@durham.ac.uk.

Abstract

We analyse stated preference data over nursing jobs collected from two different discrete choice experiments: a multi-profile case best-worst scaling experiment (BWS) prompting selection of the best and worst among alternative jobs, and a profile case BWS wherein the respondents choose the best and worst job attributes. The latter allows identification of additional utility parameters and is believed to be cognitively easier. Results suggest that respondents place greater value on pecuniary over non-pecuniary gains in the multi-profile case. There is little evidence that this discrepancy is induced by the extra cognitive burden of processing several profiles at once in the multi-profile case. We offer thoughts on other likely mechanisms.

JEL classification: C23, C25, C81, J44
Key words: discrete choice experiment, preference elicitation, rank-ordered data, latent class logit, best-worst scaling, maximum-difference model

Highlights:

- We compare preferences on nursing jobs elicited by profile and multi-profile case DCEs.
- The paper is the first to contrast the two types of DCEs using monetary and nonmonetary attributes.
- Preferences are comparable across the DCEs but only for non-monetary attributes.
- Respondents value salary gains relatively more in the multi-profile DCEs.
- The evidence suggests that this discrepancy is not due to the variation in cognitive difficulty.

1 Introduction

Discrete choice experiments (DCEs) have become a common data collection method in health economics. A recent review by de Bekker-Grob *et al.* (2012) finds that the number of published DCE studies in health economics has increased from 34 in 1990-2000 to 114 in 2001-2008. By far the most well-known type of DCE is a traditional DCE which prompts the respondent to choose her best, i.e. most preferred, profile from a set of multiple profiles differentiated by the attributes of interest. An extension to this method, the multi-profile case best-worst scaling (BWS), has become popular in recent years. By asking the respondent for both the best and worst among several differentiated profiles, it elicit more preference information with minimal additional burden (Flynn, 2010a; Flynn, 2010b).

Recently, another type of DCE has received a lot of attention in the choice modelling literature. In this alternative DCE, known as the profile case best-worst scaling (BWS), the respondent faces one hypothetical profile, and states its best and worst aspects. Within health economics, Flynn *et al.* (2007) is the primary article that kindled interest in the profile case method, and Marti (2012) discusses subsequent applications. In the present study, we provide empirical comparisons of stated preferences elicited by these two different methods: the profile case BWS and the multi-profile case BWS (Flynn 2010b). For ease of presentation and clarity, we refer to these experiments as single profile and multi-profile cases respectively.

We know of only two other papers, Potoglou *et al.* (2011) and Flynn *et al.* (2013), that study the comparability of preferences estimated using single profile and multi-profile case data. Findings in these quality-of-life studies suggest that preferences are structurally similar across methods. However, the life situations depicted in these papers include only non-monetary attributes. In contrast, our profiles include both monetary (salary) and non-monetary attributes, much as discrete alternatives of interest in a typical economic analysis. As we show below, the inclusion of monetary attributes alters the comparability of the estimated preferences substantially.

The respondents in our experiments consist of current students and recent graduates of the Bachelor of Nursing programs at two universities in New South Wales, Australia, and profiles describe typical entry-level nursing positions.[1] Each person participated in both single and multi-profile case BWS experiments during 2009-2010. A choice

[1]More information on the survey is provided in Section 2; also a detailed description is available in Kenny *et al.* (2012).

set for the multi-profile experiment includes 3 entry-level nursing jobs described by salary and 11 non-salary characteristics. By asking for both best and worst choices, we obtain a full ranking of the alternatives. A person's preferences over the job attributes are elicited to the extent that between-job variations in these attributes influence her preferences over jobs. In the single profile case experiment, each scenario is a particular nursing job described by salary and 11 non-monetary attributes set at specific levels. A respondent examines the job, and states its best and worst attributes. In effect, the person's preferences over attributes are elicited at a more primitive level. In the survey, each respondent completes 8 different single profile case scenarios followed by 8 different multi-profile case scenarios.

The users of the single profile case BWS (Flynn *et al.*, 2007; Flynn 2010a and 2010b; Marti 2012) have emphasised two advantages of this method over the multi-profile case method. First, the single profile method imposes less of a cognitive burden on the respondent. People may find it easier to understand and complete a single profile case experiment, because each scenario requires evaluation of only one hypothetical profile. This potential advantage may be especially relevant to health economics, where DCE applications often involve inherently complex profiles (e.g. medical treatments or prescription drugs). Second, the single profile case method potentially yields more information about underlying preferences. In models for the single profile case data (Marley and Louviere, 2005; Marley *et al.*, 2008), identified parameters convey whether the level of one attribute is preferred to the level of a different attribute (e.g. $1100 of salary vs excellent quality of care). In contrast, it is well-known that in models for multi-profile case data, identified parameters only convey whether a change in one attribute is preferred to a change in another attribute (e.g. increase in salary vs improvement in the quality of care).

The uptake of the single profile case BWS has been slow relative to its potential advantages. Somewhat surprisingly, this has been particularly true in health economics which is likely to benefit more than other disciplines from using a cognitively easier method (Flynn, 2010b). Most empirical studies still use multi-profile case DCEs, with little acknowledgment of the availability of the single profile case method. One contributing factor to this slow uptake may be the scarcity of empirical evidence to guide an informed choice between the two approaches.

The survey of nursing students and new graduates we analyse is well-suited to the objective of this paper. Nursing jobs are generally complex objects and many attributes are required to describe their key aspects; here we use 12 attributes in total.

In comparison, only 6 out of 148 DCE studies reviewed in de Bekker-Grob *et al.* (2012) specify more than 10 attributes. Evidence suggests that the number of attributes is the primary design dimension that influences respondents' perceived complexity of DCE tasks (Caussade *et al.*, 2005). The use of a cognitively easier method, hence, can be expected to have significant effects on respondent behaviour in the present analysis.

We specify a flexible choice model for each case, following the discrete mixture or latent class approach (Train, 2008). This approach is an attractive one for analysing two sets of preferences elicited by DCEs with different information processing requirements. Latent class models have been found to perform well in capturing key aspects of preference heterogeneity in many choice data sets (Keane and Wasi, 2012), and various heuristics in choice behaviour can be conceptualised as particular classes of preference parameters (Hensher and Greene, 2010). In the multi-profile case analysis, the flexible model's kernel is the heteroskedastic rank-ordered logit (Hausman and Ruud, 1987), and in the single profile case analysis, it is the max-diff logit (Marley and Louviere, 2005; Marley *et al.*, 2008).

Our findings suggest that in comparison with the multi-profile case method, the single profile case method elicits less noisy preferences. More importantly, the preferences are also structurally different. The key structural difference is that people value salary gains more and non-salary gains less when completing the multi-profile case experiment. To be specific, most of the utility coefficients on non-pecuniary attributes are scaled up by a similar proportion as we move from the multi-profile case to the single profile case estimates. In this regard, our results add to the existing empirical evidence from Potoglou *et al.* (2011) and Flynn *et al.* (2013). However, we also find that the utility coefficients on salary levels are scaled up by a much smaller proportion than the coefficients on non-pecuniary attributes. This result is new and it is especially important given the use of the coefficient on monetary attributes (i.e. the marginal utility of money) to estimate the valuation of non-monetary goods and attributes.

Interestingly, we find little evidence that the variation in the amount of hypothetical information to be processed drives this key discrepancy in elicited preferences. In general, only a few studies (Flynn, 2010b; Flynn *et al.*, 2013) discuss why the two methods may elicit different preference structures, and no discussion is available on specifically why the differential treatment of the pecuniary attribute may arise. We place our findings in the context of related examples from experimental economics and contingent valuation studies, and speculate on alternative driving factors some of which are amenable to further investigation using more specialised survey designs.

The remainder of this paper is structured as follows. Section 2 describes the discrete choice survey designs and estimation sample. Section 3 describes the main models to be estimated. Section 4 discusses the results and Section 5 concludes.

2 Data

We analyse discrete choice experiments collected as part of an ongoing longitudinal study of nursing job choices described more fully in Doiron *et al.* (2011) and Kenny *et al.* (2012). The data come from an online survey completed between September 2009 and September 2010. We recruited Bachelor of Nursing (BN) degree students enrolled during 2008-2010 at two large Australian universities: the University of Technology Sydney (UTS) located in a major city, and the University of New England (UNE) in a regional centre. The sample includes nursing students in each year of the 3-year BN program and new graduates (within 12 months since graduation). Nursing students include school-leavers, mature age entry and other health-care workers seeking to upgrade their qualifications. The sample covers a range of age groups, stages of household formation and exposure to nursing work.

The 526 survey respondents (100 from UNE and 426 from UTS) represent 18% of the BN enrolment at both universities during the recruitment period (19% at UNE, 18% at UTS). As discussed in Doiron *et al.* (2011), these response rates are similar to comparable cohort studies. Also, based on available demographics, there are but small differences between our cohort and all enrolled students. For more details on the sample, see Doiron *et al.* (2011).

As well as answering standard survey questions on demographics and labour market experiences, each of the 526 respondents participates in two different types of DCEs involving hypothetical entry-level nursing jobs. Each job is described in terms of salary and eleven non-salary attributes set at specific levels. The selection of attributes is based on the literature on Magnet Hospitals in the US (Naude and McCabe, 2005; Seago *et al.*, 2001), and reflect characteristics that have been shown to influence the quitting decision and job satisfaction of nurses. We use 4 different levels of salary and 2 different levels of each non-salary attribute as listed in Table 1. The salary levels reflect those found in current entry-level nursing jobs in Australia. The feedback from an earlier pilot study involving 60 students indicates that the attributes and levels are appropriate in the context of the first job as a registered nurse in Australia.

Table 1: Job attributes and associated levels

Glossary definition of attribute	Attribute name	Levels	Variable
The type of hospital where the new graduate program is located	Location	Private hospital Public hospital	Private hosp Public hosp
The number of rotations to different clinical areas	Clinical rotations	None Three	No rotation 3 rotations
Whether the new graduate program offers fulltime and part-time positions, or fulltime only	Work hours	Fulltime only Part-time or fulltime	FT hours Flex hours
The flexibility of the rostering system in accommodating requests	Rostering	Inflexible, does not allow requests Flexible, usually accommodating requests	Inflex rost Flex rost
The hospital's reputation regarding staffing levels	Staffing levels	Frequently short of staff Usually well-staffed	Short staff Well staff
The hospital's reputation regarding the workplace culture in terms of support from management and staff	Workplace culture	Unsupportive management and staff Supportive management and staff	Unsupp mgmt Supp mgt
The hospital's reputation regarding the physical work environment in terms of equipment and appearance	Physical environment	Poorly equipped and maintained facility Well equipped and maintained facility	Poor equip Well equip
The hospital's reputation regarding whether nurses are encouraged and supported in professional development and career progression	Professional development and progression	No encouragement for nurses Nurses encouraged	No encourage Encourage
The parking facilities	Parking	Limited Abundant and safe	Limit park Abund park
The hospital's reputation regarding the responsibility given to nurses, relative to their qualifications and experience	Responsibility	Too much responsibility Appropriate responsibility	Excess resp App resp
The hospital's reputation regarding the quality of patient care	Quality of care	Poor Excellent	Poor care Excell care
The gross weekly salary	Salary	$800 $950 $1,100 $1,250	Sal 800 Sal 950 Sal 1100 Sal 1250

Figure 1: Sample profile case BWS scenario (including accept-or-not DCE)

Set 8 of 8

There is a job available in a program for new graduates which has the following characteristics. Please indicate which aspect of this job you think is the **best** aspect (choose one only) and which you think is the **worst** aspect (choose one only). Please select one answer per column.

To review the features of jobs, please click here.

		Best Aspect	Worst Aspect
1. Location:	Private hospital	◯	◯
2. Clinical rotations:	Three	◯	◯
3. Work hours:	Fulltime only	◯	◯
4. Rostering:	Flexible, usually accommodating requests	◯	◯
5. Staffing levels:	Usually well-staffed	◯	◯
6. Workplace culture:	Unsupportive management and staff	◯	◯
7. Physical environment:	Poorly equipped and maintained facility	◯	◯
8. Professional development and progression:	No encouragement for nurses	◯	◯
9. Parking (The parking facilities):	Limited	◯	◯
10. Responsibility:	Appropriate responsibility	◯	◯
11. Quality of care:	Poor	◯	◯
12. Weekly Salary:	$950	◯	◯

If you were offered this job, would you take it?

◯ Yes
◯ No

Figure 2: Sample multi-profile case BWS scenario

There are jobs available in three programs for new graduates which have the following characteristics:
To review the features of jobs, please click here.

Scenario 1			
Features of Job	Job A	Job B	Job C
1. Location	Private hospital	Private hospital	Public hospital
2. Clinical rotations	Three	Three	None
3. Work hours	Part-time or fulltime	Fulltime only	Part-time or fulltime
4. Rostering	Flexible, usually accommodating requests	Inflexible, does not allow requests	Flexible, usually accommodating requests
5. Staffing levels	Usually well-staffed	Frequently short of staff	Usually well-staffed
6. Workplace culture	Supportive management and staff	Supportive management and staff	Unsupportive management and staff
7. Physical environment	Well equipped and maintained facility	Well equipped and maintained facility	Poorly equipped and maintained facility
8. Professional development and progression	Nurses encouraged	No encouragement for nurses	Nurses encouraged
9. Parking	Abundant and safe	Limited	Abundant and safe
10. Responsibility	Appropriate responsibility	Appropriate responsibility	Too much responsibility
11. Quality of care	Excellent	Poor	Poor
12. Salary	$1,250	$800	$1,100
Considering these three jobs:			
Q1. Which would you MOST like to get?	◯ Job A	◯ Job B	◯ Job C
Q2. Which would you LEAST like to get?	◯ Job A	◯ Job B	◯ Job C

The first choice experiment is the single profile case best-worst scaling (BWS). As shown in Figure 1, each scenario presents one hypothetical job and the respondent picks its best and worst aspects. The second choice experiment is the multi-profile case BWS. As shown in Figure 2, each scenario presents a choice set of three hypothetical jobs, labelled Job A, B and C, and the respondent states which job is the best and which job is the worst; all jobs are effectively ranked from most to least preferred.

Every respondent must complete 8 different scenarios of the single profile case BWS before completing another 8 scenarios of the multi-profile case BWS. This presentation sequence raises a concern that the comparability of preferences elicited by the two BWS methods is affected by fatigue. An earlier analysis of the multi-profile case data (Doiron *et al*, 2011), however, finds that the utility coefficients do not vary significantly over the 8 scenarios. Moreover, our findings on the differences in the estimates between the single and multi-profile cases do not support the wide-spread application of heuristic decision rules in the multi-profile case tasks that one may expect in the presence of respondent fatigue. We provide more details below.

We now discuss the optimality of designs underlying these two BWS experiments. The scenarios for each experiment are constructed from an initial set of 16 jobs which form a resolution 3 fractional factorial design. Initial sets for the two experiments use different resolution 3 fractions, to ensure that no multi-profile case scenario includes a job which the respondent has seen earlier in a profile case scenario.

For the single profile case experiment, the initial set of 16 jobs becomes the set of 16 scenarios. The 16-scenario set is then divided into two 8-scenario subsets, and every respondent is randomised to one of these two subsets. Street and Knox (2012) show that our design performs as well as the complete factorial design in terms of the D-criterion, when all coefficients in the standard max-diff model are equal.

For the multi-profile case experiment, each of the initial 16 jobs is augmented by a pair of new jobs to form a scenario. The two new jobs in each scenario are determined by the addition of two generators, chosen to make the resulting set of 16 scenarios D-optimal when all coefficients in the standard multinomial logit model are zero. To cover a larger proportion of the sample space, two different sets of 16 scenarios are constructed in this manner using two different resolution 3 fractions. Each 16-scenario set is then divided into two 8-scenario subsets, giving four 8-scenario subsets in total. Every respondent is randomised to one of these four subsets.

3 Model specification and selection

We begin by describing the basic notation used in the formulation of the choice models. $n = 1, \ldots, N$ denotes a respondent; $t = 1, \ldots, T$ indexes a scenario; $k = 1, \ldots, K$ indicates an attribute; $l_k = 1_k, 2_k, ..., L_k$ refers to a level of attribute k.[2] In our context, $N = 526$, $T = 8$ and $K = 12$. Each profile or job j is described by the K attributes set at specific levels. $x_{njt}^{l_k}$ is a zero-one variable which equals one if attribute k of profile j shown to respondent n in scenario t is set at level l_k.

The term "attribute-level" shall be used to describe the pair formed by an attribute and one of its possible levels. For example, when the attribute of interest is the quality of care which can be either poor or excellent, there are two possible attribute-levels: poor quality of care and excellent quality of care.

We estimate discrete mixture (latent class) models which allow utility coefficients to covary freely over a finite number of mass points (Train, 2008). These models are well-suited to our objective of comparing preferences elicited by two different methods. Keane and Wasi (2012) find that latent class logit models do well in summarising the key aspects of preference heterogeneity in many discrete choice data sets.

3.1 Models for multi-profile case data

In our multi-profile case best-worst scaling (BWS) experiment, respondents choose the best and the worst out of 3 jobs in each scenario. We thus obtain a full ranking of the jobs. These data can be modelled using the rank-ordered logit (ROL) due to Beggs *et al.* (1981). The modelling of complete rankings tend to result in smaller coefficient estimates than the modelling of first-best choices only, as if residual variances increase across preference ranks; Hausman and Ruud (1987) introduce the heteroskedastic ROL (HROL) to address this issue.

With 8 scenarios completed by each person, our data also feature a panel dimension. In the (first-best) choice modelling literature, a random parameter or "mixed" logit model (McFadden and Train, 2000) is often specified to address this dimension by capturing within-person correlations in observations as well as between-people preference heterogeneity. The same approach can be adapted for the ROL framework as demonstrated by Calfee *et al.* (2001) and Train (2008).

[2] For example, in the context of Table 1, attribute k may refer to hospital type, 1_k and 2_k being public hospital and private hospital respectively. When attribute k refers to salary, 1_k, 2_k, 3_k and 4_k are \$800, \$950, \$1,100 and \$1,250 respectively.

We analyse our multi-profile case data using an extension of HROL developed in Yoo (2012). The basic idea is to model all parameters in HROL as person-specific random parameters. In this paper, we describe the resulting model from the conventional perspective that rank heteroskedasticity arises as people are more certain about what they like more (see for example, Fok *et al*, 2012). Alternatively, Yoo (2012) motivates the use of the same model to account for stochastic misspecification of the microeconomic random utility function (McFadden, 1981). Our discussion shows that the model is an attractive tool regardless of the origin of rank heteroskedasticity.

Specifically, assume that respondent n ranks three available jobs in two statistically independent steps indexed by $r = 1, 2$. In step 1, she picks the best of the three jobs. In step 2, she eliminates her first-best from consideration, and picks the best of the other two jobs. The best job in each step is the one that provides the highest utility.

The utility she derives from job j is decomposed into a systematic component associated with attribute-levels and a random disturbance term.[3] Specifically, for $r = 1, 2$

$$U^r_{njt} = \sum_{k=1}^{K} \sum_{l_k=1_k}^{L_k} B^{l_k}_n x^{l_k}_{njt} + u^r_{njt} = \sum_{k=1}^{K} \sum_{l_k=2_k}^{L_k} \beta^{l_k}_n x^{l_k}_{njt} + u^r_{njt} = \boldsymbol{\beta}_n \cdot \mathbf{x}_{njt} + u^r_{njt} \tag{1}$$

where u^1_{njt} and u^2_{njt} are independently extreme value distributed with variances equal to $\pi^2/6$ and $\pi^2/(\sigma^2_n 6)$ respectively. $B^{l_k}_n$ measures person n's utility of having attribute-level l_k and its scale has been implicitly normalised along with the variance of u^1_{njt}. Because utility differences between jobs depend only on differences in the levels of job attributes, the utility from each attribute's first level is further normalised to 0, giving identified parameters $\beta^{l_k}_n = B^{l_k}_n - B^{1_k}_n$ for $l_k = 2_k, ..., L_k$. In consequence, $\beta^{l_k}_n > \beta^{l_l}_n$ for two different attributes k and l does not imply $B^{l_k}_n > B^{l_l}_n$. $\boldsymbol{\beta}_n$ and $\mathbf{x}_{njt}$ are vectors of identified parameters and attribute-level dummies, respectively.

Let $P_{nt}(\boldsymbol{\beta}_n, \sigma_n)$ denote the likelihood of person n's stated ranking in scenario t. Once the utility parameters $\boldsymbol{\beta}_n$ and the scale parameter σ_n are known, this likelihood takes the HROL form. For instance, if person n has ranked the three available jobs as

[3] In revealed preference applications, such random disturbance term is often associated with attributes which are known to the decision maker but unobserved by the researcher. In stated preference applications, all attributes differentiating profiles are observed, and the disturbance term can be more naturally interpreted as accounting for random fluctuations in the decision maker's state of mind. See McFadden (pp. 205-206, 1981) for a related discussion.

$1 \succ 2 \succ 3$, this likelihood becomes:

$$P_{nt}(\boldsymbol{\beta}_n, \sigma_n) = \frac{\exp(\boldsymbol{\beta}_n \cdot \mathbf{x}_{n1t})}{[\sum_{j=1}^{3} \exp(\boldsymbol{\beta}_n \cdot \mathbf{x}_{njt})]} \times \frac{\exp(\sigma_n \boldsymbol{\beta}_n \cdot \mathbf{x}_{n2t})}{[\sum_{j=2}^{3} \exp(\sigma_n \boldsymbol{\beta}_n \cdot \mathbf{x}_{njt})]} \tag{2}$$

where σ_n captures heteroskedasticity across steps in the ranking. This form of heteroskedasiticity may arise when people feel more certain about their more preferred profiles, so that their first step response depends more on systematic parts of the utility and less on random disturbances (Hausman and Ruud, 1987). σ_n would then lie in the $(0, 1)$ interval, unless person n ranks all jobs equally systematically ($\sigma_n = 1$) or picks the second-best job arbitrarily ($\sigma_n = 0$). In this view, the coefficient attenuation issue of standard ROL results from incorrectly constraining σ_n to 1.

To address the panel dimension of our data, we model $\boldsymbol{\beta}_n$ and σ_n as random parameters in a latent class framework. Specifically, we assume that there are C distinct sets or classes of utility and scale parameters. Since everyone in the same class has the same parameters, we use $\boldsymbol{\beta}_c$ and σ_c with $c = 1, ..., C$ to denote the class-specific parameters. The resulting "mixing" distribution is discrete, with η_c denoting the relative frequency of class c in the respondent population. The final likelihood of respondent n's sequence of responses over the T scenarios is specified as:

$$L_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_C; \eta_1, \ldots, \eta_C; \sigma_1, \ldots, \sigma_C) = \sum_{c=1}^{C} \eta_c \prod_{t=1}^{T} P_{nt}(\boldsymbol{\beta}_c, \sigma_c) \tag{3}$$

where $\eta_C = 1 - \sum_{c=1}^{C-1} \eta_c$. We call the model specification in equation (3) the latent class HROL (LHROL). As summarised in Table 2, LHROL can nest several other modelling approaches for rank-ordered data.

Our preferred LHROL incorporates 4 classes ($C = 4$). The estimation results are discussed in Section 4, and presented in Table 3 of the Appendix. The preferred number of classes has been chosen as in other empirical studies using latent class logit models (Greene and Hensher, 2003; Train, 2008). Specifically, the 4-class LHROL specification gave the smallest Bayesian Information Criterion (BIC) among nine alternative specifications with the number of classes varying from 2 to 10.[4]

[4]All specifications have included job-specific constants for Job A and Job B to capture potential heuristics based on labelling; interestingly Class 4, which appears to rank profiles mainly in order of salary levels, is also the only class in which these constants are significant at the 1% level.

Table 2: Nested Models in LHROL - equation (3)

Parameter restrictions	Special Cases
$C = 1$	HROL or heteroskedastic rank ordered logit (Hausman and Ruud, 1987)
$C = 1$ & $\sigma_c = 1$	ROL or rank ordered logit (Beggs *et al.*, 1981)
$C = 1$ & $\sigma_c = 0$	MNL or multinomial logit (McFadden, 1974)
$C \geq 2$ & $\sigma_c = 0, c = 1, \ldots, C$	LCL or latent class logit (Green and Hensher, 2003)
$C \geq 2$ & $\sigma_c = 1, c = 1, \ldots, C$	LCROL or latent class rank ordered logit (Train, 2008)
$C = 2$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, $\sigma_1 = 1$ & $\sigma_2 = 0$	LC-ROL or latent class rank ordered logit (Fok *et al.*, 2012)

The following summary of specification tests are based on the preferred 4-class LHROL model. The scale parameter σ_c is statistically different from 1 at the 1% level in all classes. The joint hypothesis of homogeneous ranking capabilities across all classes, that is $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, is rejected at the 4.2% level based on a Wald test statistic of 8.19. The parametric restrictions leading to LCL ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 0$) and latent ROL ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$) are overwhelmingly rejected at the 1% level using the likelihood ratio tests; the test statistics are 132.24 and 926.37 respectively.

3.2 Models for single profile case data

In each of our single profile case BWS scenarios, respondents examine *one* job described by K different attributes set at specific levels, and pick the best and the worst of these K attribute-levels. Marley and Louviere (2005) and Marley *et al.* (2008) develop alternative models to analyse the resulting data. In particular, they prove a number of choice theoretical properties of the maximum difference logit (max-diff) model, that has since been the workhorse model in empirical studies (Flynn *et al.*, 2007; Lusk and Briggeman, 2009; Lusk and Natalie, 2009; Potoglou *et al.*, 2011; Marti, 2012).

As Flynn *et al.* (2008) summarise, the max-diff model is operationalised by assuming that an observed best-worst pair is the most preferred option out of $K(K-1)$ mutually exclusive options. These $K(K-1)$ options refer to all possible best-worst pairs of attribute-levels given a profile, or job in our application. In the context of the object case BWS (Flynn, 2010b), which is like the single profile case BWS with only one possible level per attribute, Lusk and Briggeman (2009) explicitly describe it as a form of the latent dependent variable model.

We now extend the description of Lusk and Briggerman (2009) to the single profile case BWS involving several possible levels per attribute. Let respondent n's systematic utility from an attribute-level be denoted $A_n^{l_k}$.[5] Respondent n's response to a profile case scenario depends on the difference in utilities attainable from the candidate best and worst attribute-levels; specifically, she maximises the difference (hence "max-diff") between the utility from the best and the worst attribute-levels. This utility difference can be decomposed into systematic and random components.[6] In the case where attributes q and h form the candidate best-worst pair, the corresponding utility difference $D_{nt}^{\{q,h\}}$ is:

$$\begin{aligned} D_{nt}^{\{q,h\}} &= \sum_{l_q=1_q}^{L_q} \sum_{l_h=1_h}^{L_h} (A_n^{l_q} - A_n^{l_h}) x_{nt}^{l_q} x_{nt}^{l_h} + e_{nt}^{\{q,h\}} \\ &= \sum_{l_q=1_q}^{L_q} \sum_{l_h=1_h}^{L_h} (\alpha_n^{l_q} - \alpha_n^{l_h}) x_{nt}^{l_q} x_{nt}^{l_h} + e_{nt}^{\{q,h\}} \end{aligned} \tag{4}$$

where the error term $e_{nt}^{\{q,h\}}$ is independently type I extreme value distributed. The profile subscript j is dropped from attribute-level dummies $x_{njt}^{l_k}$ since only one profile is shown in each scenario.

The utility difference between any two candidate best-worst pairs will be unchanged when the same constant is added to each parameter $A_n^{l_k}$. To achieve identification, one utility parameter needs to be normalised to 0, say for the first level of the first attribute $A_n^{1_1}$. Then, each identified parameter can be defined as $\alpha_n^{l_k} = A_n^{l_k} - A_n^{1_1}$. Now $\alpha_n^{l_k} > \alpha_n^{l_l}$ for two different attributes k and l implies $A_n^{l_k} > A_n^{l_l}$; recall that a similar statement cannot be made in the context of equation (1). In this sense, with the single profile case data, we can infer more about the underlying preferences than the multi-profile case

[5] We change the notation for utility weights from $B_n^{l_k}$ to $A_n^{l_k}$ to emphasise that their scale is normalised with respect to potentially different residual variances. If the same set of primitive utility weights are applied to comparing profiles in the multi-profile case and the best-worst pairs in the single profile case, each $B_n^{l_k}$ would differ from $A_n^{l_k}$ by the same factor of proportionality.

[6] Following from footnote 3, the random component here can be interpreted as accounting for fluctuations in the decision maker's state of mind.

data.[7] Intuitively, $K-1$ more utility parameters can be identified with the single profile case experiment because it elicits preferences over attribute-levels directly, whereas the multi-profile case experiment elicits such preferences only to the extent that between-profile variations in attributes' levels affect the ranking of the profiles.

Let $F_{nt}(\boldsymbol{\alpha}_n)$ denote the likelihood of respondent n's stated best-worst pair in scenario t. Suppose that respondent n has picked q as best and h as worst. Once the identified parameters, collected as vector $\boldsymbol{\alpha}_n$, are known, this likelihood can be specified as:

$$F_{nt}(\boldsymbol{\alpha}_n) = \frac{\exp(\sum_{l_q=1_q}^{L_q} \sum_{l_h=1_h}^{L_h} (\alpha_n^{l_q} - \alpha_n^{l_h}) x_{nt}^{l_q} x_{nt}^{l_h})}{[\sum_{k=1}^{K} \sum_{l=1}^{K} \exp(\sum_{l_k=1_k}^{L_k} \sum_{l_l=1_l}^{L_l} (\alpha_n^{l_k} - \alpha_n^{l_l}) x_{nt}^{l_k} x_{nt}^{l_l})] - K} \tag{5}$$

As in the multi-profile case analysis, the utility parameters are modelled as random draws from a discrete distribution with C distinct classes, to capture inter-personal preference heterogeneity and intra-personal correlations in responses over $T = 8$ scenarios. The final likelihood of respondent n's sequence of responses is specified as a function of the relative frequency of each class c, ρ_c, and the class-specific utility parameters, $\boldsymbol{\alpha}_c$:

$$L_n(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C; \rho_1, \ldots, \rho_C) = \sum_{c=1}^{C} \rho_c \prod_{t=1}^{T} F_{nt}(\boldsymbol{\alpha}_c) \tag{6}$$

where $\rho_C = 1 - \sum_{c=1}^{C-1} \rho_c$. We call the model specification in equation (6) the latent class max-diff (LMD). LMD nests the standard max-diff model as a special case with one class, $C = 1$.

Our preferred LMD has $C = 7$ classes. As we discuss in Section 3.3, the estimated LMD parameters are transformed to make them comparable to the LHROL estimates. The transformed results are discussed in Section 4 and presented in Table 4 of the Appendix. An earlier draft of this paper provides a summary of untransformed estimates (p. 31, Yoo and Doiron, 2012). As in LHROL, the preferred number of classes has been determined by examining which of 9 alternative specifications, with C varying from 2 to 10, led to the smallest BIC.

The postulated behaviour of the max-diff model may be unrealistic in the present context, as it would have a respondent consider 132 ($= 12 \times 11$) best-worst pairs in each

[7] We emphasise that this result holds not because the single profile case data allow identifying the absolute level of the utility weight on each attribute-level; the identified parameter $\alpha_n^{l_k}$ has been explicitly normalised to represent the deviation of $A_n^{l_k}$ from $A_n^{1_1}$. This result holds because the identified parameters preserve enough information on the unidentified parameters to enable the ordinal comparisons of utility weights on different attribute-levels.

scenario. For a sensitivity check, we have also estimated a latent class variant of the sequential best-worst model (Marley and Louviere, 2005) which postulates a simpler choice behaviour: a respondent looks for the best of 12 attribute-levels, and then the worst among the other 11 attribute-levels in two statistically independent steps. This alternative model, however, has a very similar likelihood as the workhorse max-diff model, and leads to almost identical estimates; see our earlier draft (p.17, Yoo and Doiron, 2012) for further comments. Our findings are in line with Flynn *et al.* (2008) who also find the two behavioural models empirically comparable.

3.3 Normalisation convention

In LHROL (for multi-profile case data), the utility coefficient on one level of each attribute is normalised to 0. An estimated coefficient measures how much utility changes as the level of the relevant attribute changes from the omitted level to the reported level. For example, the coefficient on excellent quality of care measures the utility difference between excellent and poor qualities of care.

In LMD (for single profile case data), only the utility coefficient on the lowest salary level, $800 per week, is normalised to 0. An estimated coefficient measures the difference in utilities provided by the relevant attribute-level and the weekly salary of $800, taking the positive (negative) sign when this attribute-level is (less) preferred to $800. For example, the coefficient on excellent quality of care is positive when it provides a higher utility than the weekly salary of $800.

Our analysis focuses on comparisons of the two sets of estimates. For this purpose, the LMD coefficient estimates are transformed to represent the same information as the LHROL estimates. Specifically, the LMD coefficient on a level of each attribute is differenced with the LMD coefficient on the base level of the same attribute, where the base level refers to the omitted level in the LHROL estimation. For example, we difference the LMD coefficients on the excellent quality of care and the poor quality of care to obtain a transformed coefficient, which can be compared with the LHROL coefficient on the excellent quality of care. No transformation is required for salary, as the LMD coefficients on salary levels have already been normalised as deviations from the coefficient on $800, the omitted salary level in LHROL.

4 Main findings[8]

As discussed earlier, one advantage of the single profile case best-worst scaling (BWS) experiments over multi-profile case discrete choice experiments (DCEs) is that they yield more information on the underlying preferences. Specifically, since the single profile case BWS collects stated preferences over different attribute-levels directly, it allows identification of parameters that indicate whether a level of one attribute is preferred to a level of another attribute, instead of whether a level change in one attribute is preferred to a level change in another attribute.

However, the two methods may also elicit structurally different information. We begin this section by a discussion of hypotheses on potential differences in preference parameters across the methods. Our reading of the few studies addressing this issue suggests three broad possibilities. The first two relate to the ease of answering the profile case and favour its use, while the third calls for discretion.

First, single profile case responses may be subject to less random noise. The respondent may answer the profile case with greater certainty than the multi-profile case since the former involves consideration of one instead of several hypothetical profiles per scenario. Then, as initially envisioned by Flynn *et al* (2007), the single profile case utility coefficients would be a scaled-up version of the multi-profile case coefficients due to the smaller variance in the stochastic component. Without any structural shift in preferences, the relative magnitude of the coefficients on two different attributes would be the same across the two approaches.

Second, the single profile case task may be better understood and more attentively completed. Flynn (2010b) anticipates that the single profile case would be especially useful when the cognitive burden of processing multiple profiles is likely to be excessive. Given the complexity of jobs generally, and our use of many attributes to describe the hypothetical nursing jobs, the survey analysed in this paper provides a good example of such a situation. It is an open question as to what kind of empirical differences the varied cognitive burden may produce. A useful insight comes from a growing number of traditional DCE studies on attribute non-attendance (Cameron and DeShazo, 2008; Greene and Hensher, 2010; Hole, 2011). These studies suggest that people tend to handle the cognitive burden of a choice task by ignoring a subset of presented attributes. If an attribute is ignored in the more complex multi-profile case but taken into account

[8]All estimation results discussed in this section have been obtained using Stata 11.2/IC.

in the simpler profile case, its coefficient would have a larger magnitude relative to other coefficients in the latter case.

Third, single and multi-profile cases may elicit inherently different types of preferences. The single profile case asks respondents to state the best and worst attribute-levels of a given profile, while the multi-profile case asks respondents to trade off attributes across hypothetical profiles. Flynn (2010a; 2010b) and Flynn *et al.* (2013) suggest that tradability alters the choice context and may affect the relative magnitudes of the coefficients.[9] In our view, this is similar to effects from reference-dependent preferences in the behavioral economics literature (Kahneman, 2003).

The two studies that provide an empirical comparison of the methods (Potoglou *et al.*, 2011; Flynn *et al.*, 2013) find that the single profile case BWS and the multi-profile case DCE primarily yield the first type of discrepancy; that is, the relative magnitudes of the coefficients are preserved. In this sense, the preference structure is maintained across the two methods. Importantly, these studies do not include a monetary attribute.

The discussion of our results begins with a brief summary of findings followed by a more detailed presentation of the evidence. We start by examining to what extent the qualitative conclusions of the earlier studies hold in our data. As the earlier two studies, we find that most of the utility coefficients on non-salary attributes of the nursing jobs are scaled up by a similar proportion as we move from the multi-profile to the single profile case results. However, we also find that the utility coefficients on the monetary attribute (salary) are scaled up by a much smaller proportion; that is, respondents place a higher value on salary gains relative to improvements in other job characteristics when completing the multi-profile task.

Is the different treatment of money driven by the varied cognitive burden? More specifically, is the larger relative weight on the monetary attribute in the multi-profile case due to respondents ranking jobs mainly in order of salary to simplify multi-profile comparisons? Our results do not support this hypothesis. While one preference segment (a latent class of respondents) in our multi-profile case data is consistent with such behaviour, this segment is too small to explain the overall differences. The other classes of respondents also reduce their relative valuation of salary in the profile case task. Moreover, our analysis using an accept/reject DCE embedded at the end of each single profile case scenario further rules out any explanation tied to the cognitive burden of processing several profiles, as the estimates from this simple task are similar to those

[9]We thank an anonymous referee for the material on this point.

from the more complex multi-profile case. (More details on this survey question are provided below.)

The question then arises as to whether the two methods elicit different aspects of preferences because the multi-profile case introduces tradability. While this possibility is difficult to refute and provides a valid starting point for discussion, it does not address the question of why salary is unlike other characteristics. We speculate on more specific mechanisms by placing our findings in the context of related examples from experimental economics and stated preference analyses.

4.1 Differential treatment of salary

As mentioned above, our results are comparable to previous studies for non-monetary attributes in that most of the utility coefficients are scaled up by a similar proportion as we move from the multi-profile case to the single profile case results. The different treatment of money across the two methods is a new finding that went undetected in the earlier studies as their profiles included only non-pecuniary attributes.

Our preferred panel data models are the latent class max-diff (LMD) with 7 classes for the single profile case data and the latent class heteroskedastic rank-ordered logit (LHROL) with 4 classes for the multi-profile case data. Utility coefficients vary across classes in each model, and there is no exact correspondence between classes across the two approaches. We average utility coefficients across classes within each model using the class shares as weights, and analyse the resulting set of averages as summary statistics for the preferences elicited by each method. Figure 3 plots the average LMD coefficients against the corresponding average LHROL coefficients.[10,11] All but one of these averages are significant at the 1% level; the exception, the average LMD coefficient on public hospital (public hosp), is significant at the 6% level.[12]

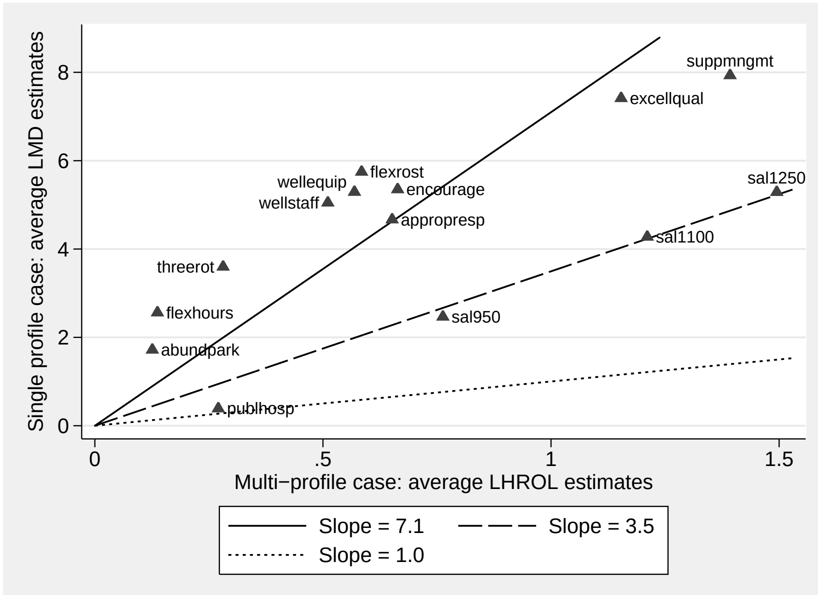
Figure 3 shows that differences between preferences elicited by the two methods cannot be entirely explained by a shift in the error variance, capturing less random noise in the profile case data. If they could, the points in this figure would be (1)

[10]Detailed estimates are available in the Appendix. Section 3.3 describes how the LMD coefficients are transformed for comparability with the LHROL coefficients.

[11]An anonymous referee pointed out that the use of a bivariate plot to summarise the estimated coefficients of two different choice models dates back to Swait and Louviere (1993).

[12]Our earlier draft (p. 21, Yoo and Doiron, 2012) presents the results from the simple max-diff and HROL models that ignore unobserved heterogeneity. The results from these simple models closely resemble those in Figure 3; the main difference is a decrease in scale which is expected since omitted preference heterogeneity increases unexplained variances (Revelt and Train, 1998).

Figure 3: single profile case vs multi-profile case coefficients



located above the dotted line with unit slope, indicating that the LMD averages are bigger than their LHROL counterparts and (2) clustered around a single steeper line with slope representing the common proportion by which the LMD coefficients are scaled up. Only the first pattern is observed.

Figure 3 also shows that the relative utility gains from two different non-salary characteristics is much more robust across data sets than relative gains involving salary and non-monetary characteristics. The bold line with the slope of 7.1 is the best fit line through the origin and the average coefficients on non-salary attributes excluding "public hospital", the characteristic that is only marginally significant in LMD. These averages are closely scattered around the bold line, whereas the average coefficients on salary levels are located far below it. In fact, the salary coefficients are very closely clustered around another line with slope equal to 3.5, suggesting that they are scaled up by roughly half as much as the non-salary coefficients.

In other words, the respondents seem to value salary gains more in relative terms when completing the multi-profile case than the single profile case. A key implication is that the rankings of average utility gains from salary and other characteristics could be reversed depending on which data set is analysed. The different treatment of salary, a monetary attribute, is especially important given the use of willingness-to-pay measures in applied work. The dollar valuation of improvement in a non-monetary attribute is usually derived from a ratio of the utility coefficient on this attribute to the marginal utility of money (the coefficient on salary). The slopes of the bold and dashed lines in Figure 3 indicate that such dollar measures can double as we move from the multi-profile to the single profile case.

4.2 Varied cognitive burden as a potential explanation

That respondents evaluate only one profile at a time has often been emphasised as a potential advantage of the single profile case BWS (Flynn *et al.*, 2007; Flynn, 2010a; Flynn, 2010b; Potoglou *et al.*, 2011; Marti, 2012). Is the greater valuation of salary in the multi-profile case driven by the cognitive burden involved in the processing of several profiles at once? We examine this possibility using two types of evidence. First, we investigate if our results are consistent with a specific form of decision heuristic derived from the growing literature on information processing strategies in DCEs (Cameron and DeShazo, 2008; Greene and Hensher, 2010; Hole, 2011; Lagarde, 2013). Second, we analyse data from a simple yes-or-no question which asks respondents whether they would accept or reject a particular job. They are asked this question for each job presented as a single profile case scenario (see Figure 1 for an example). Since this accept/reject DCE involves only one hypothetical job at a time, and each of these jobs is a single profile case scenario, we would expect the accept/reject DCE results to be comparable to the single profile case results, based on the level of cognitive difficulty.

We now turn to the first set of findings. The hypothesis is that salary influences multi-profile case responses to a greater extent because respondents rank jobs based mostly on salary as a way of handling the extra cognitive burden. Note that the same simplifying strategy cannot be applied in the single profile case, since salary must be compared with at least one other attribute to state the best and worst aspects of a job. Also, given our experimental design, each of three jobs in a multi-profile case scenario offers a distinct salary level; hence, salary can always be used to rank the three jobs.

As detailed next, we do find a preference class in the multi-profile case data which is consistent with such behaviour, and some evidence that people in this class pay more attention to non-salary attributes during the single profile case experiment. However, this class is too small to drive the overall result and we find that the relative undervaluation of salary gains in the single profile case is a wide phenomenon found in all preference classes. The finding requires more general explanations than a specific heuristic.

Following Hensher and Greene (2010) who interpret heuristics as a particular form of preferences, our analysis begins by examining whether a class with very large coefficients on salary levels and small coefficients on non-salary attributes exists in LHROL (for multi-profile case data), but not in LMD (for single profile case data).

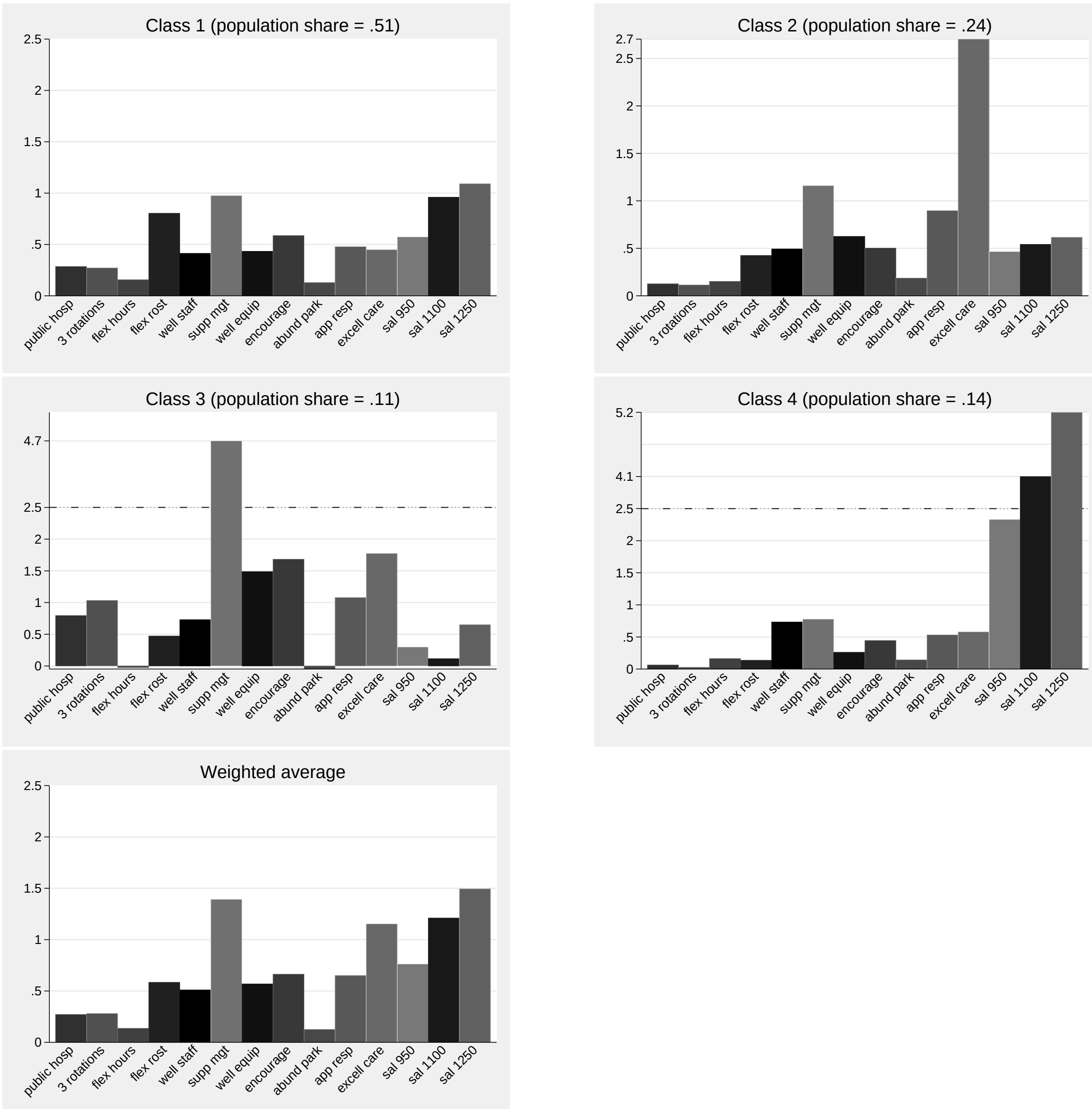
Figure 4 displays the LHROL coefficients for its four preference classes. In each panel, the horizontal axis labels attribute-levels and the vertical axis measures coefficient magnitudes.[13] Class 4, albeit with a small share of the population at 14%, indeed seems to capture individuals who rank jobs mainly in order of salary levels; we observe very big spikes in the last three columns depicting the coefficients on salary levels $950, $1100 and $1250, and much smaller bars in other columns. To aid interpretation, consider someone in Class 4 who faces two jobs. Job I pays $800 per week (smallest in our design) but has the best possible non-salary characteristics: excellent quality of care, supportive management, and so on. Job II pays a higher salary but has the worst possible non-salary characteristics. When job II pays $1250, this person has a 0.78 chance of choosing it. When job II pays $1100, she still has a 0.57 chance of choosing it, despite disadvantages in all other aspects.

Figure 5 plots the LMD coefficients for its seven preference classes.[14] The axes of each panel convey the same information as before. There is no class which shows extreme concerns for salary gains as LHROL Class 4 does. Most tellingly, Class 5 in the LMD model is the only class that places the salary increase to $1250 above any other non-pecuniary improvement. Yet, in the context of the earlier thought experiment, even this class is more likely to choose Job I than Job II paying $1250; it can be easily seen that the combined height of coefficients on non-pecuniary attributes easily exceeds the height of the coefficient on $1250. More generally, only Class 6 with a population share of 0.12 stands out in terms of likely information processing strategy. This class has small coefficients on all attributes, suggestive of a large error variance, and may capture people who expend minimal attention on the evaluation of a profile; with no obvious rule to rank attributes of the same job, some people may give responses after a very casual evaluation.

The evidence so far is consistent with the view that LHROL Class 4 captures people who use salary to simplify the multi-profile case task. In our earlier draft (p. 26, Yoo and Doiron, 2012), we complement this population-level analysis with an individual-level analysis, which further suggests that the salary-focused respondents in the multi-

[13]Table 3 in the Appendix reports the estimates.
[14]Table 4 in the Appendix presents the estimates.

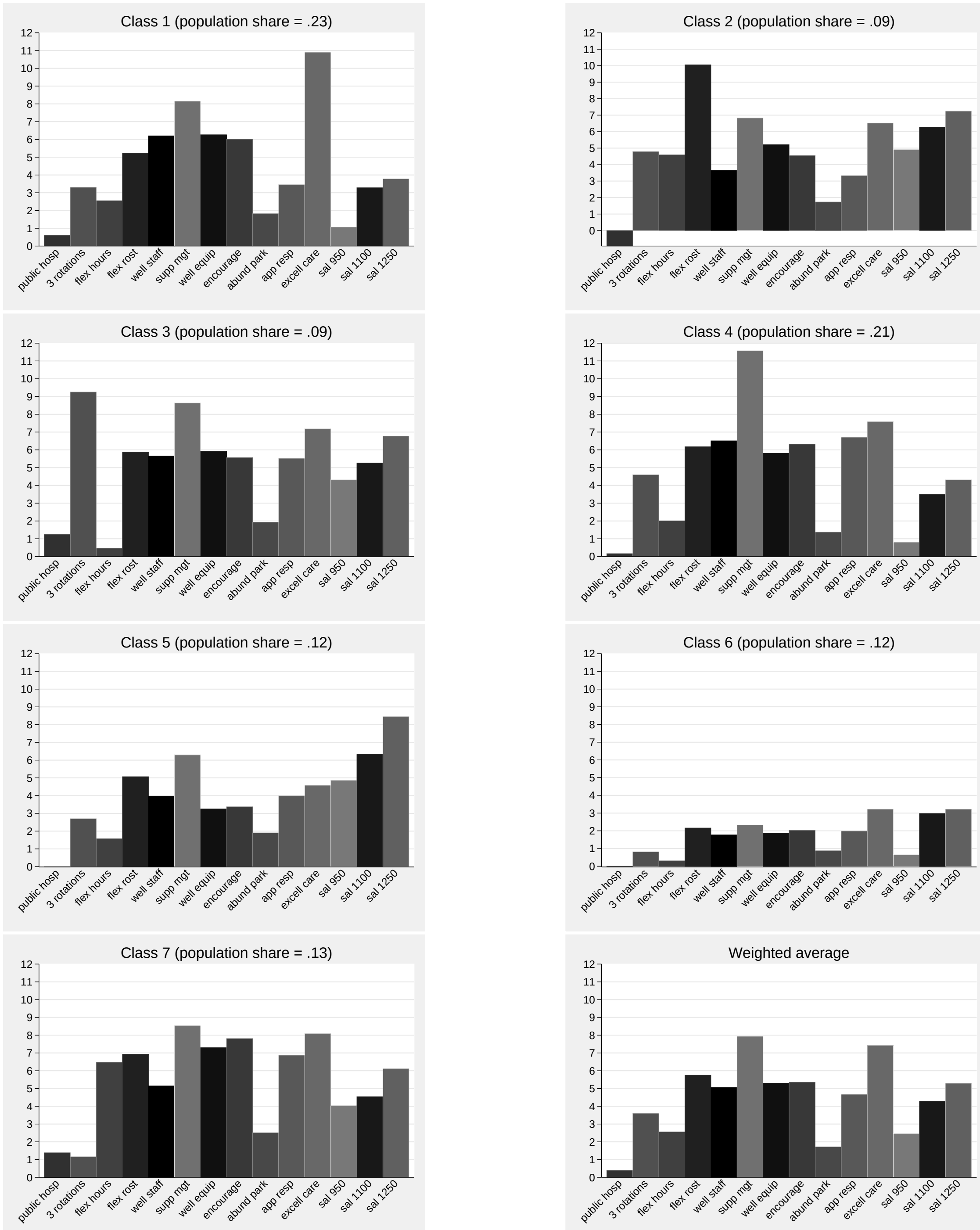
Figure 4: LHROL coefficient estimates (multi-profile case data)



profile case data have examined all attributes more attentively during the profile case experiment. Briefly, we consider someone whose multi-profile case responses are best described by LHROL Class 4, and ask which LMD classes best describe the same person's single profile case responses. Of 74 such individuals, only 18 are matched with LMD Class 6.[15]

[15]For this purpose, we compute posterior class membership probabilities (p357, Train, 2009), and match the person with the class giving the highest posterior probability in each model. Specifically,

Figure 5: LMD coefficient estimates (single profile case data)



suppose that there are C classes in total. Person n's posterior probability of membership in class c is given by $\phi_c L_{nc}/(\sum_{k=1}^{C} \phi_k L_{nk})$ where ϕ_k is the population share of class k and L_{nk} is the likelihood of observing her sequence of choices given she is in class k.

A broader comparison of preference classes in the two data sets suggests that the relative undervaluation of salary gains in the single profile case is a wide phenomenon. Figure 4 shows that LHROL Class 1 contains a majority (51%) of the population who show no striking taste for a specific attribute. This majority class values the largest salary increase (to $1250) more than any non-pecuniary gain, and the second largest increase (to $1100) more than all but one of non-pecuniary gains. By contrast, Figure 5 shows that all LMD classes, excluding Class 5 with a 12% share, value many of non-pecuniary gains more than such large salary gains. These results suggest that even those who do not exploit salary as the primary job ranking criterion value salary gains less during the single profile case experiment.

We now turn to the the second set of findings that incorporate data from a small DCE presented at the end of each single profile case scenario. Specifically, the respondent is asked to indicate if she is willing to accept the very job whose best and worst aspects she has stated. Figure 1 shows an example. For anyone who has a view on what kind of job is acceptable, this DCE should be considerably easier than the multi-profile case experiment. The person need not process two extra jobs that vary over scenarios and she only needs to consider if the job she has already evaluated provides at least her own reservation utility. Whether her view is realistic or not influences this DCE's external validity, not its comparative ease over the multi-profile case experiment.[16]

We specify a random effects (RE) logit model of the accept/reject DCE outcome, using job characteristics as regressors. The model intercept follows a normal distribution to accommodate interpersonal variations in the reservation utility or acceptability threshold. Figure 6 plots the RE logit coefficients against the average LHROL coefficients.[17] The two sets of coefficients are very similar in scale; they are closely clustered around the best fit line from the origin that has the slope of 0.8.[18]

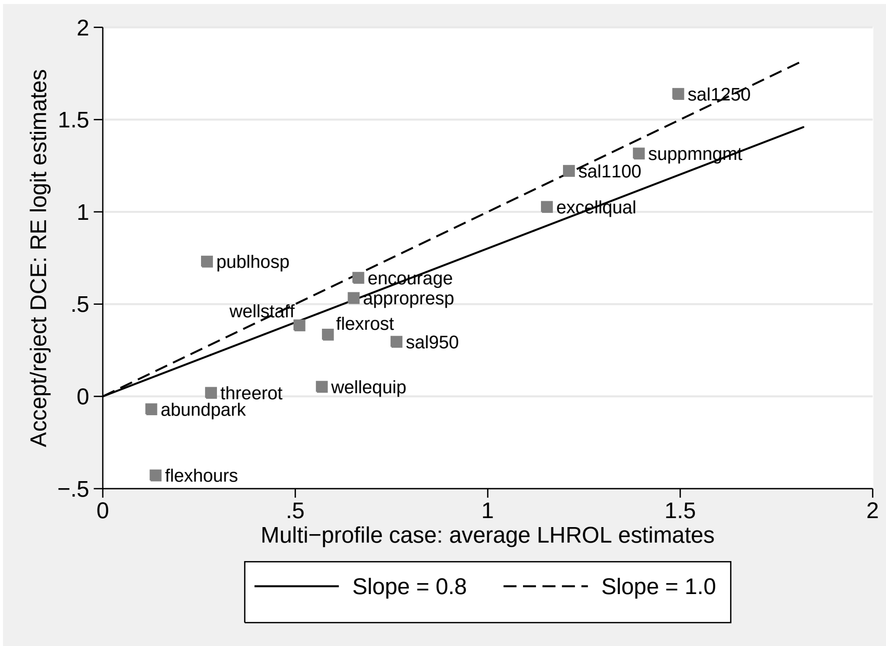
Most importantly, the RE logit and LHROL coefficients agree on how much the two largest salary gains ($1100 and $1250) are valued relative to major non-salary determinants of job choices: supportive management (supp mgt), excellent quality of care

[16]Nevertheless, the survey participants have considerable knowledge or at least strong beliefs regarding entry level jobs for nurses. Many of them have worked as nursing aides and their program includes a practicum where students get on-the-job experience. See Doiron *et.al.* (2011) for more details.

[17]In the RE logit, the intercept's standard deviation and all slope coefficients are significant at the 1% level, except those on three clinical rotations (3 rotations), well equipped (well equip) and abundant parking space (abund park).

[18]The same qualitative conclusion holds whether unobserved coefficient heterogeneity in the multi-profile case data is modelled or not. When plotting the RE logit estimates against the simpler HROL estimates, the best fit line yields a slope of 1.1.

Figure 6: accept/reject DCE vs multi-profile case coefficients



(excell care), encourage professional development (encourage) and appropriate responsibility at work (app resp). Figure 6 indicates that any pair of these coefficients has roughly the same ratio in the accept/reject DCE and the multi-profile case experiment. These findings do not support the hypothesis that it is the extra cognitive burden in the multi-profile case that led to the relative overvaluation of salary gains in Figure 3.

4.3 Other explanations

The preceding analysis suggests that justifications for greater preferences for salary gains in the multi-profile case should go beyond the processing of larger amounts of information. We close this section with a discussion of alternative explanations. Distinguishing among various behavioral hypotheses would require experiments specifically designed for the task; we also offer a few thoughts on such possible experiments as future avenues of research.

Flynn (2010b) is a rare study which explicitly speculates on why structural differences can be expected in the preferences elicited by the single and multi-profile case methods. Our reading of his argument is that the two methods invoke different choice contexts since attribute-levels are like tradable goods in the multi-profile case but not in the single profile case. What is valued in the presence of tradability could be different from what is valued in the absence of it. Flynn *et al.* (2013) investigate this hypothesis empirically in a health-related quality of life study with senior citizens, but as mentioned previously they find little evidence that the contextual variation influences the relative valuation of different attributes. Using Flynn's hypothesis to explain our findings requires consideration of why tradability affects salary and non-salary attributes to different extents.

The concept and importance of tradability may be related to reference-dependent preferences in the area of behavioral economics (Kahneman, 2003). Evidence from lab and field experiments shows that the point of reference influences the monetary valuation of a good, sometimes with surprising results. People may have evaluated the job in a single profile case scenario as the only offer in hand, whereas they are unlikely to use as the point of reference any particular one of the three job offers that a multi-profile case scenario presents. Once the job is perceived as given, the non-salary characteristics could increase in monetary value.[19] A test of this explanation could proceed with a random allocation of two contextual prompts to single profile case respondents, which would include one stating that the job is the only available offer and the other stating that it is but one of several, and comparing the preferences elicited by each treatment method with those elicited by the multi-profile case method.

An explanation more specific to our context is that the shift in preferences towards non-salary attributes in the single profile case is due to the need for direct comparisons with salary. There is a social connotation associated with certain attributes (for example, excellent quality of patient care) and several other non-salary characteristics are related to a nurse's ability to perform well in the job. Suppose for example that a single profile case hospital has the reputation for an excellent quality of patient care and pays $1250 per week. Even when the respondent regards such salary as the best aspect, she may be hesitant about stating so, to avoid revealing that she places her monetary rewards above the welfare of the patients. In the multi-profile case, because the respondent chooses best and worst jobs which differ in several aspects, the monetary value placed on non-monetary attributes is not so evident. The validity of this line of explanation could be tested by comparing single and multi-profile case experiments on objects with a price attribute and less pro-social elements, for example different knee injury treatment options as in Bryan *et al.* (2000).

[19]This is an example of endowment effects. For a recent contribution to this literature, see Ericson and Fuster (2011).

Finally we note that our results are consistent with results from comparative studies on willingness-to-pay (WTP) derived using multi-profile DCEs and different contingent valuation methods. As summarised in Ryan and Watson (2009), a more direct elicitation format tends to generate a smaller WTP for an intervention. Translated in terms of our application, this means that a more direct elicitation format leads to a lower dollar valuation of improvements in non-salary characteristics. The multi-profile case, which prompts trading off attributes across profiles, is indeed a more direct way to elicit how much salary gains a person is willing to forgo for a better non-salary characteristic; the single profile case collects more primitive information on what attribute-level is preferred to another, which we then use to infer trade-offs. This aspect of the two cases can also be seen from the need to transform the profile case parameters for comparability with the multi-profile case parameters (see the earlier discussion in Section 3.3). To our best knowledge, there are no behavioral explanations to date for these latter findings.

5 Conclusion

We have analysed stated preference data from two different discrete choice experiments (DCEs): multi-profile case best-worst scaling (BWS) which, like traditional DCEs, involves choices over several profiles, and single profile case BWS which involves choices over attributes of a given profile. In our application, a profile is an entry-level nursing job. That respondents need to process only one profile at a time, and may thus understand single profile case tasks better, has been often advanced as an advantage of the single profile BWS method. Also, the ability to identify additional utility parameters can make the single profile case BWS a profitable alternative to multi-profile DCEs. For example, in relation to our application, suppose that hospital managers are considering how best to allocate a fixed budget to the design of new nursing jobs meant to attract nurses away from non-nursing jobs. A relevant multi-profile DCE may be hard to design, because jobs in different occupations are best described by different attributes. A single profile case BWS experiment would provide useful inputs by allowing inference of attribute-levels which are more preferred than others, thereby highlighting key features an attractive nursing job needs to possess.

We find that when restricting attention to non-pecuniary attributes, the relative valuation of different non-pecuniary gains remains fairly stable across the two cases. In contrast, the elicited preferences over gains in pecuniary (salary) and non-pecuniary attributes vary substantially, with the multi-profile case analysis indicating much stronger preferences for pecuniary gains. Our results, however, show that the differential treatment of salary requires explanations which go beyond different amounts of information the respondents need to process in the two experiments. An economic analysis is mostly concerned with objects which involve both pecuniary and non-pecuniary aspects. In this respect, our analysis suggests that if the single profile case BWS is to become a broadly accepted method, empirical evidence from more specialised studies is needed to inform why profile and multi-profile DCEs may elicit different preferences for the two distinct aspects. It is hoped that our earlier discussion will provide a basis for future research in this direction.

References

Beggs S, Cardell S, Hausman J. Assessing the potential demand for electric cars. Journal of Econometrics 1981; 16; 1-19.

de Bekker-Grob E, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Economics 2012; 21; 145-172.

Bryan S, Gold L, Sheldon R, Buxton M. Preference measurement using conjoint methods: an empirical investigation of reliability. Health Economics 2000; 9; 385-395.

Cameron T, DeShazo J. Differential attention to attributes in utility-theoretic choice models. Journal of Choice Modelling 2008; 3; 73-115.

Caussade S, Ortuzar J, Rizzi L, Hensher D. Assessing the influence of design dimensions on stated choice experiment estimates. Transportation Research Part B 2005; 39; 621-640.

Dockery A, Barns A. Who'd be a nurse? some evidence on career choice in Australia. Australian Bulletin of Labour 2005; 31; 4; 350-383.

Doiron D, Hall J, Kenny P, Street D. Job preferences of students and new graduates in nursing. 2011; CHERE Working Paper 2011/2.

Ericson K, Fuster A. Expectations as endowments: evidence on reference-dependent preferences from exchange and valuation experiments. Quarterly Journal of Economics 2011; 126; 1879-1907.

Fok D, Paap R, Van Dijk B. A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. Journal of Applied Econometrics 2012; 27; 831-846.

Flynn T, Louviere J, Peters T, Coast J. Best-worst scaling: what it can do for health care research and how to do it. Journal of Health Economics 2007; 26; 171-189.

Flynn T, Louviere J, Peters T, Coast J. Estimating preferences for a dermatology consultation using best-worst scaling: comparison of various methods of analysis. BMC Medical Research Methodology 2008; 8; 76.

Flynn T. Using conjoint analysis and choice experiments to estimate QALY values. Pharmacoeconomics 2010a; 28; 711-722.

Flynn T. Valuing citizen and patient preference in health: recent developments in three types of best-worst scaling. Expert Review of Pharmacoeconomics and Outcomes Research 2010b; 10; 259-267.

Flynn T, Peters T, Coast J. Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data. Journal of Choice Modelling 2013; 6; 34-43.

Greene W, Hensher D. A latent class model for discrete choice analysis: contrasts with mixed logit. Transportation Research Part B 2003; 37; 681-698.

Hausman J, Ruud P. Specifying and testing econometric models for rank-ordered data. Journal of Econometrics 1987; 34; 83-104.

Hensher D, Greene W. Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification. Empirical Economics 2010; 39; 413-426.

Hole A. A discrete choice model with endogenous attribute attendance. Economics Letters 2011; 110; 203-205.

Kahneman D. A psychological perspective on economics. American Economic Review; 2003; 93; 162-168.

Keane M, Wasi N. Comparing alternative models of heterogeneity in consumer choice behavior. Journal of Applied Econometrics; 2012; DOI: 10.1002/jae.2304.

Kenny P, Doiron D, Hall J, Street D, Milton-Wildey K, Parmenter G. The training and job decisions of nurses - the first year of a longitudinal study investigating nurse recruitment and retention. 2012; CHERE Working Paper 2012/02.

Lagarde M. Investigating attribute non-attendance and its consequences in choice experiments with latent class models. Health Economics; 2013; 22; 554-567.

Lancsar E, Louviere J. Estimating Individual level discrete choice models and welfare measures using best worst choice experiments and sequential best worst MNL. 2008; CenSoC Working Paper 08-003.

Lloyd A. Threats to the estimation of benefit: are preference elicitation methods accurate?. Health Economics 2003; 12; 393-402.

Lusk J, Briggeman B. Food values. American Journal of Agricultural Economics 2009; 91; 184-196.

Lusk J, Natalie P. Consumer preferences for amount and type of fat in ground beef. Journal of Agricultural and Applied Econometrics 2009; 41; 75-90.

Marley A, Louviere J. Some probabilistic models of best, worst, and best-worst choices. Journal of Mathematical Psychology; 2005; 49; 464-480.

Marley A, Flynn T, Louviere J. Probabilistic models of set-dependent and attribute-level best-worst choice. Journal of Mathematical Psychology; 2008; 52; 281-296.

Marti J. A best-worst scaling survey of adolescents' level of concern for health and non-health consequences of smoking. Social Science & Medicine 2012; 75; 87-97.

McFadden D 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka P (Ed), Frontiers in Econometrics. Academic Press: New York; 1974. p. 105-142.

McFadden D 1981. Econometric models of probabilistic choice. In: Manski C, McFadden D (Eds), Structural analysis of discrete data: with econometric applications. MIT Press: Cambridge; 1981. p. 197-270.

McFadden D, Train K. Mixed MNL models for discrete response. Journal of Applied Econometrics 2000; 15; 447-470.

Naude M, McCabe R. Magnet hospital research pilot project conducted in hospitals in Western Australia. Contemporary Nurse 2005; 20; 38-55.

Potoglou D, Burge P, Flynn T, Netten A, Malley J, Forder J, Brazier J. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. Social Sciences and Medicine 2011; 72; 1717-1727.

Revelt D, Train K. Mixed logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics 1998; 80; 647-657.

Ryan M, Watson V. Comparing welfare estimates from payment card contingent valuation and discrete choice experiments. Health Economics 2009; 18; 389-401.

Seago J, Ash M, Spetz J, Coffman J, Grumbach K. Hospital registered nurse shortages: Environmental, patient, and institutional predictors. Health Services Research 2001; 36; 831-852.

Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research 1993; 30; 305-314.

Train K. EM Algorithms for Nonparametric Estimation of Mixing Distributions. Journal of Choice Modelling 2008; 1; 40-69.

Train K. Discrete choice methods with simulation. Cambridge University Press: New York; 2009.

Yoo H, Doiron D. The use of alternative preference elicitation methods in complex discrete choice experiments. UNSW Australian School of Business Research Paper No. 2012-16; 2012.

Yoo H. The perceived unreliability of rank-ordered data: an econometric origin and implications. UNSW Australian School of Business Research Paper No. 2012-46; 2012.

Appendix

Table 3: LHROL estimation results (multi-profile case)

Variable	Class 1	Class 2	Class 3	Class 4	Weighted Average
Sal 950	0.574***	0.466***	0.300	2.332***	0.763***
	(0.084)	(0.150)	(0.336)	(0.351)	(0.080)
Sal 1100	0.961***	0.541***	0.116	4.141***	1.211***
	(0.098)	(0.158)	(0.277)	(0.472)	(0.098)
Sal 1250	1.093***	0.617***	0.653*	5.151***	1.495***
	(0.106)	(0.176)	(0.377)	(0.502)	(0.112)
Supp mgt	0.976***	1.161***	4.702***	0.778***	1.393***
	(0.067)	(0.126)	(0.965)	(0.202)	(0.106)
Excell care	0.450***	2.704***	1.775***	0.580***	1.154***
	(0.069)	(0.229)	(0.457)	(0.130)	(0.076)
App resp	0.479***	0.897***	1.080***	0.534***	0.652***
	(0.063)	(0.122)	(0.344)	(0.153)	(0.055)
Flex rost	0.804***	0.425***	0.473***	0.138	0.585***
	(0.059)	(0.108)	(0.160)	(0.149)	(0.043)
Encourage	0.585***	0.503***	1.683***	0.445***	0.664***
	(0.059)	(0.104)	(0.425)	(0.138)	(0.056)
Well equip	0.433***	0.626***	1.488***	0.262*	0.569***
	(0.055)	(0.108)	(0.492)	(0.148)	(0.063)
Well staff	0.413***	0.493***	0.730**	0.732***	0.511***
	(0.054)	(0.106)	(0.284)	(0.145)	(0.047)
Public hosp	0.285***	0.127	0.795**	0.064	0.271***
	(0.055)	(0.103)	(0.391)	(0.142)	(0.054)
3 rotations	0.270***	0.115	1.035**	0.029	0.281***
	(0.056)	(0.107)	(0.415)	(0.150)	(0.058)
Flex hours	0.157***	0.152	-0.027	0.164	0.137***
	(0.052)	(0.101)	(0.129)	(0.127)	(0.038)
Abund park	0.129**	0.186*	-0.049	0.145	0.126***
	(0.052)	(0.098)	(0.144)	(0.140)	(0.040)
Job B Cst	0.123*	0.054	-0.135	0.395**	0.117**
	(0.066)	(0.113)	(0.214)	(0.173)	(0.049)
Job A Cst	0.050	-0.291*	-0.167	0.511***	0.009
	(0.062)	(0.154)	(0.217)	(0.169)	(0.055)
σ	0.508***	0.504***	0.954***	0.675***	0.578***
	(0.045)	(0.061)	(0.230)	(0.100)	(0.038)
Class	0.513***	0.241***	0.107***	0.139***	
share	(0.034)	(0.031)	(0.020)	(0.019)	
Number of respondents		526	Log likelihood		-5706.48
Number of observations		21040	BIC		11832.74

BIC refers to the Bayesian information criterion. Asymptotic standard errors are in parenthesis. *** indicates that the parameter is significantly different from zero at the 1% level, ** at 5% and * at 10%.

Table 4: Transformed LMD estimation results (single profile case)

Variable	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Weighted Average
Sal 950	1.077*	4.913***	4.329***	0.810	4.868***	0.658**	4.034***	2.468***
	(0.622)	(0.863)	(0.888)	(0.814)	(0.476)	(0.333)	(1.074)	(0.277)
Sal 1100	3.287***	6.278***	5.265***	3.496***	6.320***	2.981***	4.543***	4.278***
	(0.512)	(0.680)	(0.777)	(0.569)	(0.478)	(0.342)	(1.045)	(0.217)
Sal 1250	3.785***	7.245***	6.773***	4.314***	8.454***	3.216***	6.110***	5.290***
	(0.511)	(0.694)	(0.726)	(0.556)	(0.914)	(0.346)	(0.968)	(0.225)
Supp mgt	8.151***	6.830***	8.642***	11.581***	6.294***	2.328***	8.542***	7.934***
	(0.381)	(0.568)	(0.611)	(0.418)	(0.416)	(0.289)	(0.926)	(0.199)
Excell care	10.904***	6.520***	7.187***	7.592***	4.582***	3.222***	8.087***	7.415***
	(0.424)	(0.574)	(0.602)	(0.428)	(0.414)	(0.247)	(0.929)	(0.194)
App resp	3.455***	3.332***	5.522***	6.709***	3.987***	1.987***	6.872***	4.670***
	(0.474)	(0.721)	(0.675)	(0.436)	(0.406)	(0.281)	(0.940)	(0.198)
Flex rost	5.230***	10.054***	5.874***	6.178***	5.070***	2.162***	6.921***	5.749***
	(0.425)	(0.576)	(0.701)	(0.426)	(0.420)	(0.286)	(0.863)	(0.189)
Encourage	6.010***	4.548***	5.561***	6.321***	3.371***	2.025***	7.807***	5.352***
	(0.403)	(0.648)	(0.647)	(0.447)	(0.449)	(0.274)	(0.940)	(0.189)
Well equip	6.266***	5.212***	5.915***	5.810***	3.255***	1.872***	7.298***	5.294***
	(0.393)	(0.624)	(0.677)	(0.489)	(0.459)	(0.285)	(0.852)	(0.192)
Well staff	6.204***	3.647***	5.651***	6.510***	3.963***	1.772***	5.148***	5.051***
	(0.414)	(0.901)	(0.663)	(0.426)	(0.440)	(0.292)	(0.920)	(0.192)
Public hosp	0.610	-0.945	1.246*	0.164	-0.017	-0.039	1.388*	0.392*
	(0.485)	(0.659)	(0.737)	(0.505)	(0.494)	(0.265)	(0.813)	(0.208)
3 rotations	3.305***	4.792***	9.254***	4.600***	2.700***	0.815***	1.162	3.600***
	(0.463)	(0.776)	(0.596)	(0.507)	(0.539)	(0.284)	(0.836)	(0.210)
Flex hours	2.557***	4.597***	0.468	2.011***	1.573***	0.309	6.475***	2.564***
	(0.481)	(0.617)	(0.872)	(0.622)	(0.501)	(0.273)	(1.041)	(0.217)
Abund park	1.826***	1.732**	1.928***	1.370**	1.904***	0.880***	2.513***	1.721***
	(0.494)	(0.759)	(0.736)	(0.597)	(0.461)	(0.270)	(0.821)	(0.214)
Class	0.234***	0.086***	0.093***	0.210***	0.122***	0.119***	0.135***	
share	(0.024)	(0.016)	(0.017)	(0.023)	(0.018)	(0.017)	(0.025)	
Number of respondents	526					Log likelihood	-12261.59	
Number of observations	555456					BIC	25657.208	

The coefficients are transformed for an easier comparison with the results from the LHROL model; specifically, coefficients are differenced with respect to the base level for each attribute. BIC refers to the Bayesian information criterion. Asymptotic standard errors are in parenthesis. *** indicates that the parameter is significantly different from zero at the 1% level, ** at 5% and * at 10%.