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

Hong il Yoo
Denise Doiron

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

Hong il Yoo,
School of Economics, University of New South Wales, Australia.
Email: h.yoo@unsw.edu.au

Denise Doiron,
School of Economics, University of New South Wales, Australia.
Email: d.doiron@unsw.edu.au

Abstract

We analyse stated preference data over nursing jobs collected from two leading types of best-worst discrete choice experiments (DCEs): a traditional DCE involving choice over alternative jobs (BWL) and a newly-developed DCE where respondents choose best and worst job attributes (BWT). 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 traditional DCE. Rather than caused by the use of heuristics in BWL, we find that respondents find it difficult and/or are reluctant to directly compare money with other attributes in BWT.

Key Words: complex discrete choice experiments, preference elicitation, rank-ordered data, latent classes, heteroskedastic logits, maximum-difference models

---

Acknowledgments: This work was made possible by Discovery Project Grant DP0881205 from the Australian Research Council. We are especially grateful to the project research team: Jane Hall, Debbie Street and Patsy Kenny. We also wish to thank participants at the AHES conference 2011, and workshops at UNSW. We are especially thankful to Agne Suziedelyte for outstanding research assistance.

Corresponding author: Hong il Yoo, School of Economics, Level 4, ASB Bldg., University of New South Wales, 2052, Australia. Email: h.yoo@unsw.edu.au

1 Introduction

In many contexts, revealed preferences data with information on choices made by consumers under actual market settings may not be suitable, as the choice situation of interest may contain a currently non-existent alternative or the amount of independent variation in attributes tends to be limited. In response, many empirical researchers have analysed stated preferences data obtained through discrete choice experiments (DCEs) in which respondents are asked to state their choices under hypothetical settings.

In a traditional DCE, the decision maker is asked to rank-order, either partially or fully, hypothetical alternatives described by attributes set at different levels. For example, in our context the respondent ranks 3 jobs, each job described by a vector of attributes. The various levels of these attributes (e.g. salary) differentiate the jobs from one another. The respondent's preferences over attributes are elicited indirectly to the extent that variations in the attribute levels influence her rankings of available alternatives. In the remainder of the paper, we use best-worst alternative (BWL) to denote a DCE in which respondents are asked to state their most and least preferred alternatives.

A small but growing number of recent studies have administered another type of DCE (Flynn *et al.*, 2007; Lusk and Briggerman, 2009; Lusk and Natalie, 2009; Potoglou *et al.*, 2011) in which the decision maker is presented with one hypothetical alternative described by attributes set at specific levels, and asked to select the best and worst attributes. To continue with our example, the respondent is presented with one job described by a vector of attributes set at specific levels, and asked to choose her most and least preferred attribute. In contrast to the traditional DCE, the preferences over attributes are elicited directly, albeit partially. We refer to this experiment as best-worst attribute-level (BWT) to emphasise the choice task completed by the respondent.[1]

The proponents of BWT (Flynn *et al.*, 2007) argue that it allows the collection of richer and more systematic information on the decision maker's preferences over attributes. The information is richer since choices directly reflect comparisons over attribute levels; in other words, the identified parameters represent normalised utility weights on attribute *levels*. In models describing a traditional DCE choice task, the identified parameters can be only used to examine whether a *change* in the levels of one attribute is preferred to that of another attribute. Furthermore, in BWT the respondents face a less cognitively demanding choice task as they are presented with one instead of several hypothetical alternatives at a time. The resulting information is possibly more accurate as respondents may become less prone to making mistakes or using heuristics. This can be especially important in complex choice tasks such as ranking jobs based on a large set of characteristics. It is worth noting though that BWT has its own drawbacks. In particular, some comparisons may become harder to make when the context provided by a well-defined set of alternatives is removed. At the least, models based on these data should recognise that when left unspecified, the alternatives used by respondents in making their choices are likely to vary based on unobservables and hence preference heterogeneity can be magnified.

The debate over which type of DCE may be the better preference elicitation method has been mainly one-sided as a vast majority of studies, particularly in economics, are based on data obtained by traditional DCEs. There is rarely any discussion of potential gains and dangers from using alternative methods such as BWT.

The slow uptake of BWT in the literature may be partly explained by the paucity of

[1] This type of experiment is also known as best-worst scaling in the literature.

studies which compare preferences estimated from the two approaches. In contrast, there have been several studies comparing estimates of willingness-to-pay obtained by direct survey and traditional DCEs. The findings generally show that preference estimates differ significantly across methods possibly because respondents adopt different answering strategies (Lloyd, 2003). Likewise, preferences elicited by the BWL and BWT methods may be structurally different, as these methods vary in terms of choice tasks as well as the amount of information that needs to be processed to complete those tasks. To aid an informed choice between the two methods, it is important to accumulate evidence on the comparability of the estimated preferences and examine whether any discrepancies emerge in different contexts.

The primary objective of this paper is to compare preferences estimated from BWL and BWT data on the same group of respondents. In the process, we formulate a generalised rank-ordered logit (ROL) specification for panel data, which incorporates heterogeneity in utility parameters as in mixed logit (McFadden and Train, 2000; Train, 2008) and allows for more general forms of heterogeneity in ranking capabilities than recent extensions of ROL (e.g. Fok *et al.*, 2011). We also discuss how various discrete choice models for traditional DCE data can be applied to BWT data. This is an important extension to the current literature as these data have been almost exclusively analysed via the max-diff model. (We describe the max-diff model in detail below.)

The analysis is based on a unique dataset involving nursing students and new graduates of nursing programs collected from two universities in New South Wales (see Doiron *et al.*, 2011). In the experiment, each respondent completes 8 different BWT tasks and 8 different BWL tasks. An alternative corresponds to a hypothetical nursing job described by specific levels of salary and eleven other job characteristics. Since jobs are inherently complex, many attributes are needed to describe their key aspects and the use of a method that is cognitively easier for respondents may have pronounced effects on their responding strategies. Hence, this survey is well-suited to the objectives of the paper.

To the best of our knowledge, Potoglou *et al.* (2011) is the only other study to compare these two approaches empirically. The results in that paper favour the use of BWT. An alternative in the Potoglou *et al.* study is a hypothetical living situation, described by social care related qualities of life. The authors report that once the utility parameters estimated from the BWT data are transformed to make them comparable to those estimated from the traditional DCE data, most of the BWT estimates are larger in magnitude by roughly the same proportion. In other words, preferences elicited by BWT seem to be a less noisy version of those elicited by traditional DCEs. This result combined with the added information provided by BWT supports the use of the richer and less cognitively demanding method.

In this paper, we find that preferences elicited by BWT and BWL methods exhibit an important systematic difference that cannot be ascribed to a reduced unexplained variance in the BWT data alone. Specifically, as we move from the BWL estimates to the BWT estimates, most of the utility coefficients associated with non-monetary attributes are scaled up by a similar proportion. This is consistent with the results in Potoglou *et al.* (2011). However, the estimates associated with salary are scaled up by a much smaller proportion. In other words, respondents tend to place a higher value on an increase in salary relative to an improvement in another job aspect when completing the BWL task than the BWT task. We note that all attributes in Potoglou *et al.* are non-pecuniary; hence, this comparison could not be made.

The different treatment of monetary characteristics has important implications for empirical studies based on discrete choice methods. Researchers routinely derive a dollar

measure of the welfare change from a policy intervention or other change affecting a non-pecuniary attribute by using the ratio of the coefficient associated with that attribute and the estimate attached to a monetary attribute. The latter represents the marginal utility of money. Our findings suggest that conclusions from such studies may be sensitive to the choice between BWT and traditional DCE methods. If a pecuniary improvement is relatively more valued with BWL as in our study, the implied dollar value of welfare change becomes smaller in magnitude *ceteris paribus*.

We examine two alternative explanations for the systematic difference between the two sets of preference estimates. First, as the recent literature on attribute non-attendance suggests (Cameron and DeShazo 2008; Hensher and Greene, 2010; Hole, 2011), some respondents may have heuristically ranked alternatives mainly in order of salary levels to simplify the complex BWL choice task. The same heuristic decision rule cannot be applied to the BWT choice task since the respondent is forced to make comparisons across different attributes. Second, some respondents may inflate the value of non-salary job characteristics when completing the BWT task due to the incentive to hide true preferences or the difficulty with directly comparing salary and non-salary characteristics. For example, out of ethical considerations, respondents may be reluctant to state that a specific salary level, say $1100 per week, is better than providing excellent quality of care to patients. Additionally, it may be difficult to determine whether the salary level is preferred to the quality of care in the absence of a well-defined alternative job.

We find empirical evidence that is more in line with the latter explanation. Given the patterns of preference heterogeneity and the magnitude of the preference weights, it is difficult to ascribe the differences in the two sets of preference parameters to the application of simple heuristics. In the BWL data, a majority of respondents are found to prefer the largest possible salary gain in our survey to an improvement in any other attribute. However, they are willing to trade off this salary gain for a simultaneous improvement in two or more attributes. Only a small fraction of respondents are estimated to rank-order heuristically (as defined in Hensher and Greene, 2010), placing the largest salary gain above a simultaneous improvement in all other characteristics. The population share of these potential users of heuristics is too small to explain the overall preference for salary in BWL.

Further support for the second explanation is provided by an analysis using the data from a yes-or-no question at the end of each BWT task asking respondents whether they are willing to accept the hypothetical job. Since the alternative job is not specified, each respondent is presumed to compare the stated BWT alternative to her outside option. This yes-or-no task is much less complex than the BWL task and hence there is less incentive to use simple heuristic choice rules. Yet the utility coefficients estimated from these data are more similar to the BWL estimates than the BWT estimates, suggesting that the difficulty with directly comparing salary to other attributes under the BWT design is a more important driver of the pattern of differences discussed above.

In brief, our results suggest that the verdict in favor of BWT in the existing literature may be too rosy and that the impact of the two approaches on estimated preferences is more complex. In our results, BWT generates a fairly neutral effect in terms of relative weights for non-monetary attributes in that coefficients are increased roughly proportionately presumably due to a reduction in the scale or variance of random utility. But we also find that in BWT, the relative weights between the monetary and non-monetary characteristics of a good are distorted. Although tentative, further analysis of our data suggests that the differences in the estimates derived from the two approaches may reflect a reluctance or difficulty in directly comparing monetary and other attributes

in BWT rather than the adoption of a simple decision rule (such as selection based on the monetary attribute only) in the traditional DCE. The disproportional reduction in the utility weights on monetary attributes in BWT is especially important given the use of these weights in the calculation of willingness-to-pay measures.

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). The data come from an online survey completed between September 2009 and September 2010. Our sample was recruited from the Bachelor of Nursing (BN) degree students enrolled during 2008-2010 at two large Australian universities: one located in a major city, the University of Technology Sydney, and the other located in a regional centre, the University of New England. The sample consists of nursing students in each year of the 3-year program and new graduates (within 12 months of completing their university degree). 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 choice of attributes is based on the nursing literature, in particular the studies on 'magnet hospitals' in the US (Naude and McCabe, 2005; Seago *et al.*, 2001). They reflect characteristics that have been shown to matter in 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 Appendix Table I. The levels of the attributes, in particular salary, 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.

In the first choice experiment, the best-worst attribute level (BWT), respondents examine a scenario representing one hypothetical job and pick its best and worst aspects(see Figure 1). The second choice experiment is a traditional DCE in which respondents examine a scenario of three hypothetical jobs, labelled Job A, B and C (see Figure 2). Respondents state which they think is the best job and which they think is the worst job; all jobs are effectively ranked from most to least preferred. We call this experiment the best-worst alternative (BWL).

Every respondent must complete the BWT task for eight different scenarios before completing the BWL task for another eight different scenarios. This sequence of presentations raises a concern that the comparability of preferences elicited from the two experiments is affected by fatigue. An earlier analysis of the BWL data (Doiron *et al*, 2011) finds that the utility coefficients do not vary significantly across the eight scenarios within the BWL experiment. Moreover, our findings on the differences in the estimates across experiments do not support the wide-spread application of simple heuristics in the BWL tasks as one would expect in the case of respondent fatigue. We provide more details below.

We now discuss the optimality of designs underlying these two choice experiments. The scenarios for each experiment are constructed from an initial set of 16 jobs which

form a resolution 3 fractional factorial design. For the BWT experiment, this set becomes the set of scenarios and 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; see Street and Knox (2012) for detailed derivations. This set of 16 jobs is then divided into two subsets or versions of 8 scenarios and each respondent is randomised to one of the resulting 2 versions.

For the BWL experiment, the other two jobs in each scenario are determined by the addition of two generators, chosen so that the resulting set of 16 scenarios of size 3 is D-optimal when all coefficients in the standard multinomial logit model are zero. Two sets of 16 scenarios are constructed using two different resolution 3 fractions so that a larger proportion of the sample space is covered. These fractions differ from the fraction used in the BWT experiment design to ensure that no respondent has already examined one of the jobs in a BWL scenario. Each set is divided into two subsets or versions of 8 scenarios and each respondent is randomised to one of the resulting 4 versions.

3 Model specification and selection

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

We use the term "attribute-level" 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 or latent class models which allow random coefficients to covary freely over a finite number of mass points. The choice of a discrete instead of continuous mixing distribution is mainly driven by practical considerations. Since each job in our survey is described by a large number of attributes, it is difficult to estimate a continuous mixture model without heavily restricting the correlations among utility coefficients on different attribute-levels (Chapter 6, Train, 2009). In addition, our analysis focuses on comparisons of preferences elicited by two different methods and it is somewhat easier to interpret the representation of preference heterogeneity in the latent class framework.

3.1 Models for best-worst alternative (BWL) data

In the BWL component of our survey, respondents effectively rank-order all three jobs as they select the best and the worst out of three. We thus obtain a data structure which is usually analysed via the rank-ordered logit model (ROL) due to Beggs *et al.* (1981). A key feature of rank-ordered data is that estimated utility coefficients tend to become attenuated as the ROL likelihood is specified to include more complete rankings of alternatives. To address this issue, Hausman and Ruud (1987) introduce the heteroskedastic ROL (HROL) that parameterises shifts in error variances across ranks, while Fok *et al.* (2011) formulate a latent class ROL in which some respondents are assumed to assign completely arbitrary rankings to less preferred alternatives. Both of these approaches

conceptualise that some respondents may be more capable of ranking alternatives that are strongly liked.

Our BWL data set also has a panel dimension as it contains eight observations for each respondent. For repeated data on choices of best alternative, it is now common to specify a random parameter or "mixed" logit model (McFadden and Train, 2000) to capture correlations across observations within each respondent and preference heterogeneity across respondents. The same modelling approach can be adapted for the ROL framework as demonstrated by Calfee *et al.* (2001) and Train (2008).

For the analysis of the BWL data, we estimate a new extension of the ROL as discussed in Yoo (2012). The main idea is to model all parameters in Hausman and Ruud (1987)'s HROL as individual-specific random coefficients. The resulting specification incorporates both heterogeneity in preferences over attributes and heterogeneity in ranking capabilities, thereby nesting the above modelling approaches as special cases.

Specifically, assume that respondent n rank-orders three available jobs in two statistically independent steps indexed by $r = 1, 2$. In step one, she chooses the best of three jobs, and in step 2, she chooses the best from the two remaining jobs after excluding the job picked in step 1. The best job in each step is the one that provides the highest utility. Following the standard random-utility framework (McFadden, 1973), the utility derived from a job is decomposed into a systematic component associated with attribute-levels and a random disturbance term. Specifically, for $r = 1, 2$

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

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. $B^{l_k}_n$ is the systematic utility from attribute-level l_k derived by respondent n and its scale has been implicitly normalised along with the variance of u^1_{njt}. Because only changes in utility in response to changes in the levels of attributes matter when choosing among alternatives, the utility from the first level of each attribute is set to 0 defininining the normalised 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 parameters and attribute-level dummies, respectively.

Let $P_{nt}(\boldsymbol{\beta}_n, \sigma_n)$ denote the probability of the rank-ordering stated by respondent n on choice occasion t. Once the utility parameters $\boldsymbol{\beta}_n$ and the scale parameter σ_n are known, this probability can be specified in the HROL form. For instance, if the respondent ranks the three alternative jobs as $(1 \succ 2 \succ 3)$, the probability becomes:

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

where σ_n captures heteroskedasticity across steps in the ranking. According to Hausman and Ruud (1987), this form of heteroskedasiticity may arise as respondents feel more certain about their most preferred alternative and hence rank the least preferred alternatives less systematically. σ_n is expected to lie in the $(0, 1)$ interval unless the respondent ranks all jobs equally systematically ($\sigma_n = 1$) or chooses the second-best job completely arbitrarily ($\sigma_n = 0$). In this context, the coefficient attenuation issue in rank-ordered data mentioned above results from incorrectly restricting σ_n to 1.

To allow for preference heterogeneity across respondents, we model $\boldsymbol{\beta}_n$ and σ_n as random coefficients in a latent class framework. Specifically, we assume that there are C

distinct sets or classes of utility and scale parameters. Since each individual in a class has identical parameters, we use $\boldsymbol{\beta}_c$ and σ_c with $c = 1, ..., C$ to denote these parameters. The resulting "mixing" distribution is discrete and η_c is used to denote the relative frequency of each class c in the respondent population. The likelihood of respondent n's sequence of responses over the T scenarios can be written as:

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

where $\eta_C = 1 - \sum_{c=1}^{C-1} \eta_c$. Note that the number of classes, C, must be specified prior to estimation to achieve identification. We call the model specification in equation (3) the latent class HROL (LHROL). As summarised in Table I, several modelling approaches for rank-ordered data can be shown to be nested in LHROL.

With $C = 1$, our model reduces to Hausman and Ruud (1987)'s heteroskedastic rank-ordered logit which in turn reduces to the usual rank-ordered logit when there is no rank heteroskedasticity. Also with $C = 1$ and $\sigma_c = 0$, our model is equivalent to the mutlinomial logit that uses the data on the most preferred alternative only. With $C = 2$, the LHROL nests Fok *et al.* (2011)'s latent class rank ordered logit as a special case occurring when $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, $\sigma_1 = 1$ and $\sigma_2 = 0$. In this context, σ_c can be interpreted as a ranking capability parameter and some respondents are assumed to rank the less preferred alternatives arbitrarily because they are not familiar enough with the choice situation to provide detailed rankings. Our specification is more general in that respondents are assumed to possess different levels of ranking capabilities rather than either full or no capability; *i.e.* each σ_c in our specification is a free parameter.

With $C \geq 2$, the LHROL is equivalent to latent class logit (LCL) or mixed logit with a discrete mixing distribution (Greene and Hensher, 2003) when $\sigma_c = 0$ for all classes so that only the first best choice data are used to estimate $\boldsymbol{\beta}_c$, and reduces to latent class ROL (Train, 2008) when $\sigma_c = 1$ for all classes so that all available data are used without accounting for potential heteroskedasticity across ranks. Maintaining that LHROL is true, the former modelling approach leads to a less efficient estimator as fewer data are used and the latter leads to an inconsistent estimator in the presence of rank heteroskedasticity.

Estimation results for LHROL are discussed in the next section and detailed in Appendix Table II. Our preferred specification is estimated with four classes. In choosing the number of classes, we follow the literature on latent class logit models (Greene and Hensher, 2003; Shen, 2009; Train, 2008; Hess *et al.*, 2011). Initially, we estimated nine LHROL specifications with the number of classes varying from 2 to 10 and found that the Bayesian Information Criterion (BIC) was minimised with the use of four classes. Note that all specifications have included alternative-specific constants (ASCs) for Job A and Job B to capture potential heuristics based on labelling; interestingly Class 4, which appears to rank alternatives mainly in order of salary levels, is also the only class in which these constants are significant at the 1% level.

The following results from specification tests are based on the preferred model (LHROL with 4 classes). 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 using the Wald test statistic computed as 8.19. The parametric restrictions associated with 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 relevant test statistics are 132.24 and 926.37 respectively.

As a sensitivity check, we compared the estimation results of LCL, Train (2008)'s latent class ROL and LHROL. In general, the class share weighted average of each coefficient is similar in magnitude across LCL and LHROL but more precisely estimated in the latter. The most notable exception relates to the alternative specific constants; the average LHROL estimates of the ASCs are less precise and much smaller than the corresponding LCL estimates.[2] This result is favourable for LHROL given that the job labelling in our experiment is arbitrary. The average estimates from Train (2008)'s latent class ROL have smaller magnitudes than the corresponding LHROL estimates, most of them by 20%, qualitatively suggesting the presence of rank heteroskedasticity.

3.2 Models for best-worst attribute-level (BWT) data

In a BWT experiment, respondents examine one alternative described by K different attributes set at specific levels, and select the best and the worst of these K attribute-levels. In empirical applications (Flynn *et al.*, 2007; Lusk and Briggeman, 2009; Lusk and Natalie, 2009; Potoglous *et al.*, 2011), the observed choice is modelled as the most preferred option out of $K(K-1)$ mutually exclusive best-worst pairs of attribute-levels.

The maximum difference (max-diff) model provides a behavioural foundation for this modelling approach. Lusk and Briggeman (2009) explain this model in a random utility framework in the case of a BWT experiment with one level per each attribute. We generalise their discussion to an experiment involving a potentially different number of levels for each attribute.

Respondent n derives a systematic utility from each attribute-level, denoted by $A_n^{l_k}$.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 a potentially different error variance. The respondent bases her choice on the difference in utilities attainable from the candidate best and worst attribute-levels; specifically, the respondent maximises the difference between the utility from the best and the worst attribute-levels. This utility difference can be decomposed into systematic and random components. Suppose that 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 alternative subscript j is dropped from attribute-level dummies $x_{njt}^{l_k}$ since only one alternative is shown in each choice occasion.

The systematic difference in utility differences across 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 is 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

[2]The mean LHROL ASC for Job A is 0.009 (0.16) while the mean LCL ASC for Job A is 0.039 (0.50). The corresponding figures for Job B are 0.117 (2.38) and 0.231 (3.68) respectively. The numbers in parentheses are asymptotic t statistics.

that a similar statement cannot be made in the context of equation (1). In this sense, with the BWT data, we can infer more about the underlying preferences than the BWL data. More intuitively, $K-1$ more utility parameters can be identified with the BWT data because the respondent directly compares attribute levels whereas in BWL data, alternatives are chosen based on *changes* in attribute levels only.

Let $F_{nt}(\boldsymbol{\alpha}_n)$ denote the probability of the best-worst pair actually chosen on occasion t by respondent n. Suppose that the respondent picks q as best and h as worst. Once the identified parameters collected in the vector $\boldsymbol{\alpha}_n$ are known, the associated probability can be written 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 BWL analysis, the utility parameters are modelled as random draws from a discrete distribution with C distinct classes to capture preference heterogeneity across respondents and correlations across the 8 observations from the same respondent. The 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 utility parameters for that class, $\boldsymbol{\alpha}_c$:

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

where $\rho_C = 1-\sum_{c=1}^{C-1}\rho_c$. As in LHROL, the number of classes, C, must be determined *a priori*.

We call the model specification in equation (6) the latent class max-diff (LMD) and discuss the results obtained with $C=7$ in the next section. When the number of classes is varied from 2 through 10, BIC is lowest at 7 classes. LMD reduces to the standard max-diff model when only one class is specified.

The max-diff specification is the primary workhorse in the current literature on BWT. However, when there are 12 attributes as in our survey, the decision-making process would have respondents consider 144 best-worst pairs when making their choices. Hence, the behavioural assumptions underlying this modelling approach seem unrealistic when the number of attributes is larger than 3 or 4.

As a sensitivity check, we consider an alternative specification that describes a simpler decision making process. The structure of BWT data is equivalent to that of data obtained by a BWL experiment presenting more than three alternatives per scenario. Accordingly, models for incomplete rankings from a traditional DCE can be reinterpreted to analyse the BWT data. Let Q_{nt}^k denote the utility derived by respondent n from the attribute level k describing the job shown in scenario t:

$$Q_{nt}^k = \sum_{l_k=1_k}^{L_k}\Gamma_n^{l_k}x_{nt}^{l_k} + v_{nt}^k = \sum_{l_k=1_k}^{L_k}\gamma_n^{l_k}x_{nt}^{l_k} + v_{nt}^k \tag{7}$$

where $\Gamma_n^{l_k}$ is the systematic utility associated with this attribute-level and v_{nt}^k is i.i.d. type I extreme value distributed. Note that the error term, v_{nt}^k, is now associated with a single attribute level. In ranking attribute levels, only differences in Q_{nt}^k matter and the sytematic utility from one attribute-level must be normalised. $\gamma_n^{l_k}$ represents the normalised utility weights with one attribute level, say $\Gamma_n^{1_1}$, set to 0. As for the max-diff model, all other parameters can be identified. Equation (7) is identical to the random utility model motivating multinomial logit with alternative-specific constants and no

other explanatory variables. The probability of observing q as the best attribute takes the usual MNL expression: $\exp(\sum_{l_q=1_q}^{L_q} \gamma_n^{l_q} x_{nt}^{l_q}) / [\sum_{k=1}^{K} \exp(\sum_{l_k=1_k}^{L_k} \gamma_n^{l_k} x_{nt}^{l_k})]$. One can also derive the probability of a best-worst pair $\{q, h\}$ as the joint probability over all complete rankings yielding the choice of q as best and h as worst. This is the approach in Van Ophem *et al.* (1999) and can be seen as a variant of the rank ordered logit for incomplete rankings.

The incomplete rankings probabilities with known best-worst choices based on equation (7) become algebraically cumbsersome when K is large as in our survey, and we consider a sequential best-worst logit (SBWL) due to Lanscar and Louviere (2008) instead. As the name suggests, it is assumed that the respondent sequentially chooses the best out of K attribute-levels and the worst out of the remaining $K-1$ attribute-levels. These two steps are also assumed to be statistically independent. The probability of observing q and h as the best-worst pair is (mis)specified as:

$$G_{nt}(\boldsymbol{\gamma}_n) = \frac{\exp(\sum_{l_q=1_q}^{L_q} \gamma_n^{l_q} x_{nt}^{l_q})}{[\sum_{k=1}^{K} \exp(\sum_{l_k=1_k}^{L_k} \gamma_n^{l_k} x_{nt}^{l_k})]} \times \frac{\exp(\sum_{l_h=1_h}^{L_h} -\gamma_n^{l_h} x_{nt}^{l_h})}{[\sum_{k \neq q} \exp(\sum_{l_k=1_k}^{L_k} -\gamma_n^{l_k} x_{nt}^{l_k})]} \tag{8}$$

Since the extreme value distribution is asymmetric, the true probability that h is chosen as worst does not equal the second ratio on the right hand side of equation (8) but a more algebraically involved expression as derived in the Appendix to Fok *et al.* (2011). The difference between these probabilities, however, tends to be small empirically as the distribution is only slightly asymmetric.

We have estimated a latent class SBWL similar to LMD by modelling γ_n as draws from a discrete distribution. The resulting estimates are negligibly different from the LMD estimates for any number of classes varying from 1 to 10. To see why, notice that the SBWL probability in equation (8) is algebraically similar to the max-diff probability in equation (4) and in fact, they become identical when $-K$ is added to the denominator of equation (8) and $\sum_{k \neq q} \exp(\sum_{l_k=1_k}^{L_k} -\gamma_n^{l_k} x_{nt}^{l_k})$ is replaced with $\sum_{l=1}^{K} \exp(\sum_{l_l=1_k}^{L_l} -\gamma_n^{l_l} x_{nt}^{l_l})$. The results from the latent class SBWL model are available upon request.

3.3 Normalisation convention

In LHROL, 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, only the utility coefficient on the lowest level of salary ($800 per week) is normalised to 0. An estimated coefficient measures the difference in utilities provided by the relevant attribute-level and a salary of $800 per week, and takes a 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 gives a higher systematic utility than a salary of $800 per week.

Our main 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 comparable to the LHROL coefficient on the

excellent quality of care. This transformation is not required for salary, as the LMD coefficients have been already normalised relative to $800, which is the omitted salary level in LHROL.

4 Main findings

The proponents of best-worst attribute-level (BWT) argue that the key advantage of this method over traditional DCEs (including BWL) is that BWT can be used to obtain richer information on the underlying preferences. Specifically, because BWT collects stated preferences over different attribute-levels directly, it allows estimation of models in which a sufficient number of utility parameters are identified to infer whether a level of one attribute is preferred to a level of another attribute. However, BWT and traditional DCEs may also elicit structurally different information. Individuals may respond to particular survey instruments by varying the amount of attention they expend as well as the extent to which they state their true preferences. For instance, elsewhere in the literature on stated preference methods, it is well known that the willingness to pay for an improvement in an attribute as estimated by direct surveys and traditional DCEs tend to disagree due partly to the use of different responding strategies (Lloyd, 2003). We begin by examining whether utility parameters estimated using BWT and BWL data lead to broadly similar inferences about the relative importance of improvements in different attributes.

Our preferred panel data models, latent class max-diff (LMD) with seven classes for the BWT data and latent class heteroskedastic rank-ordered logit (LHROL) with four classes for the BWL data, assign different utility parameters for different latent classes and there is no exact correspondence between classes across the two models. We average utility parameter estimates across classes within each model using the estimated class shares as weights and analyse the resulting set of averages as summary statistics for the preferences estimated from each data set. Also, as noted in the previous subsection, the LMD estimates are transformed to be comparable to the LHROL estimates.

The two sets of estimates may be said to be broadly similar when most of the average LMD estimates are larger in magnitude than the corresponding LHROL estimates by roughly the same proportion. Then a strong case can be made in favour of using BWT over BWL because preferences elicited by BWT can be said to be both richer in the amount of preference information provided and less noisy. In standard non-linear discrete choice models, all identified coefficients are scaled up by the same proportion following a decrease in the error variance, keeping the relative magnitude of any two coefficients unchanged. Potoglou *et al.* (2011) find this type of similarity in their social quality of life survey using multinomial logit and max-diff models with fixed utility parameters.[3]

Figure 3 plots the average LMD coefficients against the corresponding average LHROL coefficients. (Detailed estimates are presented in Appendix Tables 2 and 3.) All but one of the averages are significant at the 1% level, the exception being the average LMD coefficient on public hospital (public hosp) which is significant at the 6% level only.

For comparison, Figure 4 plots corresponding estimates from fixed coefficient versions of the max-diff and HROL models; all estimates are significant at the 1% level. The average random parameter estimates in Figure 3 resemble closely the fixed parameter estimates in Figure 4, the main difference being an increase in scale. This is not surprising

[3]The panel dimension of their data is addressed by specifying independently normally distributed individual-specific intercepts instead of random utility coefficients on attribute-levels.

as the explicit modelling of preference heterogeneity tends to reduce error variances (Revelt and Train, 1998).

The broad conclusion to be drawn from Figure 3 is that the difference between preferences elicited by the two methods cannot be explained simply by a smaller amount of random variation in the BWT data. If it could, most of the points in Figure 3 would be (1) located above the dotted line with the unit slope, indicating that the LMD averages are bigger than their LHROL counterparts and (2) clustered around a steeper line, the slope of which would represent the common proportion by which the max-diff coefficients are scaled up. Only the first pattern can be clearly observed in this figure.

Figure 3 shows that the relative importance of two different non-salary characteristics is much more robust across data sets than that of a salary level and a non-salary characteristic. The bold line with the slope of 7.1 in Figure 3 is the best fit line through the origin and the averages associated with non-salary attributes excluding "public hospital", the characteristic that is only marginally significant in BWT data. These averages are closely scattered around the bold line, though they do not line up exactly, whereas the average coefficients on salary levels are located far below it suggesting that the latter set of averages are scaled up by a much smaller proportion than the former. In fact the salary coefficients are very closely clustered around a line with slope equal to 3.5, approximately one half of the proportion used to scale the non-salary attributes. In words, the respondents as a group seem to value salary gains more in relative terms when completing BWL than BWT. One implication of the patterns in Figure 3 is that the rankings of average utility gains from salary and other characteristics could be reversed depending on which data set is analysed.

We consider whether the estimation results are consistent with the view that respondents complete the BWT task more systematically as is often argued by proponents of this approach. The recent literature on the problem of attribute non-attendance (Cameron and DeShazo, 2008; Greene and Hensher, 2010; Hole, 2011) provides a useful link between this argument and the structural differences found above. These studies suggest that in a traditional DCE, respondents may make choices heuristically based on a subset of available attributes and ignore the rest to minimise the cognitive burden. For example, in the present context, a respondent may rank alternatives in order of salary levels alone in BWL. Similar heuristics may not be easily applied in BWT since the person needs to compare salary with at least one other attribute to state both best and worst aspects of a job. The presence of respondents who use heuristics in BWL may make individuals appear as a group to value salary gains more in BWL than in BWT.

Do the differences in the two sets of estimated preferences plausibly originate from the use of salary-based heuristics in BWL? To answer this question, we follow Hensher and Greene (2010)'s interpretation of heuristics as a particular class of preferences, and are interested in whether a class with very large coefficients on salary gains and small coefficients on changes in other attributes is estimated to exist in LHROL but not in LMD.

Figure 5 displays the LHROL coefficient estimates for its four latent classes. In each of the four panels, the horizontal axis labels attribute-levels and the vertical axis measures the magnitude of the coefficients. (The estimates can be found in Appendix Table II.) Of particular interest is Class 4 which is estimated to represent 14% of decision makers. This class represents respondents who rank alternatives mainly in the increasing order of salary levels, paying only minimal attention to variations in other characteristics. Graphically, we observe big spikes in the last three columns corresponding to the coefficients on salary changes from \$800 to \$950, \$1100 and \$1250, respectively and much smaller bars in other

columns. Statistically, six out of the eleven coefficients on non-salary attributes are insignificant at the 5% level; the other five coefficients, along with the three coefficients on salary gains, are significant at the 1% level.

To quantify the implications of these estimates, suppose that a decision maker in Class 4 chooses between the following two jobs. Job I pays the lowest level of salary ($800 per week) but has the best possible combination for the other eleven characteristics: excellent quality of care, an appropriate level of responsibility, supportive management and so forth. Job II pays a higher salary but has the worst possible levels for all other characteristics. When job II pays $1250 the decision maker has a 0.78 chance of choosing it, and when it pays $1100 she is still more likely to choose it, with an estimated probability of 0.57, despite the disadvantage in all other aspects. She is less likely to choose job II only when it offers the smallest possible salary gain to $950, in which case the predicted probability is 0.18. If this class of preferences reflects the use of heuristics in answering a complex BWL question, instead of indicating that many decision makers consider salary as a much more important aspect of a job than other characteristics, then we would expect these individuals to change their behaviour when faced with the BWT experiment.

Figure 6 plots the transformed LMD coefficient estimates for each of its seven classes. (The estimates are presented in Appendix Table III.) All class shares and all but a handful of untransformed coefficients have been precisely estimated at the 1% level, even though the specification involves 181 parameters, benefitting from a small amount of random variations in the BWT data.

The estimates suggest that although there is a substantial amount of preference heterogeneity across classes, no class has an extreme preference for an improvement in a particular attribute, including salary. Class 5, which accounts for 12% of the respondent population, exhibits the strongest preferences for salary gains but this class does not disregard other improvements to the same extent as Class 4 in LHROL does; for example the change from "unsupportive" to "supportive" management is estimated to result in a similar utility gain as a salary increase from $850 to $1100. In terms of classes exhibiting signs of heuristics, the most likely type is Class 6, also with a 12% population share. The magnitudes of all utility gains are much smaller here than in any other class, suggesting that Class 6 tends to state their preferences much less systematically than others. One possible interpretation is that Class 6 is similar to Class 4 in LHROL, and represents people who try to expend minimal attention; as there is no simple rule to rank attribute-levels within the same alternative, these respondents may state preferences after casually examining the presented information and appear to make choices arbitrarily.

Since both datasets share the same set of respondents, we can compare the behaviour by individuals across the two types of experiments. Specifically, we compare posterior class membership probabilities across BWL and BWT data for respondents whose BWL choices are best described by Class 4 in LHROL, the class highlighted as potentially using heuristics based on salary. Posterior membership probabilities refer to the probabilities of each respondent belonging to different classes, conditional on her observed sequence of choices. Specifically, suppose that there are C classes in total. The 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 the agent's sequence of choices given she is in class k.

The average of the largest posterior probability over all 526 respondents is 0.89 for LHROL and 0.90 for LMD, suggesting that both models do well in distinguishing different classes of preferences. In LHROL, Class 4 gives the largest posterior membership prob-

ability for 74 respondents (the average probability across these 74 individuals is 0.91). In which LMD classes are these 74 people most likely to belong? Interestingly, Class 5 and Class 6 in LMD are associated with the highest and second highest average posterior probabilities, 0.28 and 0.24 respectively.[4] At a disaggregated level, Class 5 and Class 6 give the highest posterior probability for 54% (or 40 respondents) of this subsample, 22 (30%) in Class 5 and 18 (24%) in Class 6 respectively. In other words, a nontrivial fraction of respondents whose BWL choices are best summarised by extreme concerns for salary gains (LHROL Class 4) make BWT choices consistent with the minimal expenditure of attention (LMD Class 6). Yet, a big majority of them make BWT choices implying preferences for salary gains which are strong but not extreme (LMD Class 5) or much weaker (other five LMD classes).

The evidence so far is consistent with the hypothesis that BWT may induce some respondents to consider non-pecuniary attributes which they would ignore in the more complex BWL experiment. However, the relative undervaluation of salary gains in BWT cannot be ascribed to the lessened use of heuristics alone. The group identified as potentially using heuristics in BWL (Class 4 of LHROL) is estimated to represent only 14% of the respondent population. The LHROL estimates in Figure 5 indicate that a majority (51%) of the population belongs to Class 1 which trades off different attributes fairly without extreme preferences for any subset of attributes. This segment of the population still gains the highest utility from the largest possible increase in salary to $1250. By constrast, in all classes of LMD in Figure 6 excluding Class 5, improvements of several non-salary attributes lead to utility gains similar to or higher than what is generated by the largest salary increase. The systematic undervaluation of salary in BWT requires alternative explanations that can be applied to a wider segment of respondents.

What else may be driving the observed patterns of results? We conjecture two hypotheses related to the fundamental format of the BWT task. First, the comparison of a salary level with a non-salary characteristic as in BWT may be much more difficult to make than that of a salary gain with an improvement in a non-salary attribute as in BWL, because there is no clearly good or bad level of salary. Whether a salary level is good or not may depend on the perceived salary available in other jobs. Unlike BWL which explicitly specifies the job choice set, BWT presents only one job per scenario and leaves the alternatives unspecified. In contrast, with the exception of hospital type (public vs private) and perhaps the number of clinical rotations (three vs none), non-salary attributes in our survey have inherently good and bad levels, for instance the staffing level being either set at well-staffed or short of staff. This feature of the survey design may facilitate comparisons among non-salary attributes, and make their good levels more likely to be chosen as best compared to salary levels which may be perceived as neither good nor bad. Figures 3 and 4 show that the relative magnitudes of coefficients on non-salary attributes with good and bad levels are much more consistent across the two sets of data than those of coefficients on salary and non-salary attributes without good and bad levels.

Second, respondents may tend to understate their true preferences for salary levels over other attribute-levels in BWT out of moral concerns. It is well known that when asked to state willingness to pay for environmental preservation, survey respondents tend to quote an amount larger than those estimated through a traditional DCE in which the environmental quality and income (or costs) are used to describe hypothetical alternatives. The reverse problem may occur in BWT. To illustrate, suppose that the respondent is presented with a job at a hospital which provides an excellent quality of

[4]The third highest average posterior probability is much lower at 0.13.

care to patients and the highest possible level of salary, $1250. Even when the respondent regards salary as the best aspect, she may be hesitant about stating so to avoid revealing directly that she places her own monetary benefit above the welfare of the patients. This type of consideration is less likely to influence her response in BWL where she states preferences over alternatives varying in the levels of salary and other characteristics simultaneously.

To investigate this issue further, we turn to a third source of data, namely the accept-or-not question at the end of each BWT scenario. As shown in Figure 1, respondents are asked to state whether they are willing to accept each hypothetical job after choosing its best and worst aspects. This binary task may be cognitively easier and less subject to salary-based heuristics than BWL because each respondent can compare the hypothetical job she has already seen against her own opt-out option. At the same time, she may evaluate the relative attractiveness of the job's salary level more easily because now she is choosing between the stated job and a well-defined alternative from her perspective.[5] Also, there is less incentive to downplay preferences for salary over other attributes since the alternative is not specified.

We are interested in whether the relative magnitudes of salary and non-salary coefficients estimated using these binary choice data are more similar to those estimated from the BWT data or from the BWL data. The former is more likely if the use of salary-based heuristics in BWL mainly generates differences between the two sets of results while the latter is more likely if the design of the BWT task is the primary driver. Of course it is possible that preferences elicited by the accept-or-not task are very different from those elicited by either BWT or BWL, as the latter two methods assume participation in the nursing workforce while the accept-or-not task allows the respondent's opt-out option to be a non-nursing job.

The accept-or-not decision is modelled as a random effects (RE) logit using job characteristics as explanatory variables. The intercept is assumed to follow a normal distribution to account for variations in the unobserved opt-out option across respondents.[6] The RE logit estimates are plotted against the average LHROL estimates in Figure 7 and the fixed coefficient HROL estimates in Figure 8. All coefficients and the standard deviation of the intercept in the RE logit model are significant at the 1% level, except for those associated with three clinical rotations (3 rotations), well equipped (well equip) and abundant parking space (abund park). As in Figures 3 and 4, the dotted line has a slope of one. Even though the BWL task is much more complex than the accept-or-not task, the estimated coefficients are very similar in scale whether preference heterogeneity in the BWL data is modelled or not; the slope of the best fit line through the origin is 0.8 in Figure 7 and 1.1 in Figure 8.

We do not observe that respondents place systematically more value on an increase in salary relative to improvements in other characteristics when answering the BWL experiment compared to the accept-or-not type of question. Moreover, the RE logit and LHROL estimates agree on the magnitudes of the coefficients on the two largest salary gains (from $800 to $1100 and $1250) relative to the coefficients on major non-salary determinants of the job choice in BWL (supportive management, excellent quality

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

[6]More general finite and continuous mixture models that allow for random slope coefficients as well as a random intercept have been estimated but only very imprecisely; the data are not rich enough to allow the disentangling of a random shift in the intercept from random variations in other coefficients. Result are available upon request.

of care, encourage professional development and appropriate responsibility at work); graphically, the points corresponding to these coefficients closely line up along the bold line, indicating that the ratio of any pair of these coefficients remains roughly constant across the two sets of estimates.

Simply asking respondents to state whether they are willing to accept the hypothetical job presented for BWT, without providing any additional information, leads to similar estimated preferences as those estimated from the BWL data. Based on these comparisons, we conclude again that the main structural difference between the BWT data and the BWL data, in terms of how salary changes are valued relative to variations in other attributes, are more likely to result from the format of the BWT task than the wide use of heuristics in BWL.

To conclude this discussion, we return to the second advantage of the BWT approach namely the provision of richer preference data. Recall that in our max-diff model, each identified coefficient measures the difference in utilities from the associated attribute-level and the salary level of $800. Table II reports the weighted averages across 7 classes of the untransformed LMD coefficients after ranking them in decreasing order.

In this case, the estimation results suggest that on average, there is a close connection between what we can learn from the two sets of data. For instance, an analysis of the BWL data shows that a change in the quality of care is a major determinant of job choices while the estimates using the BWT data suggest that this is because its good level (excellent) is the most preferred and bad level (poor) is the least preferred attribute-level. A similar conclusion holds across the board; if the change in one attribute is the k^{th} most important determinant of nursing job choices according to the BWL analysis, then its good and bad levels tend to be the k^{th} most and k^{th} least preferred attribute-levels according to the BWT analysis. Of course this could be very different in another context. Respondents could have valued the excellent quality of care less than appropriate responsibility and at the same time shown a lot more aversion to poor quality of care than excessive responsibility.

One motivation for our survey is the poor retention rates of nurses in Australia. An extensive analysis using the BWL data concludes that improvement in salary, managerial support for professional development and the quality of care at the institution is most likely to make a job more attractive than other nursing jobs (Doiron *et al.*, 2011). A natural follow-up question is: what are the key characteristics of an attractive nursing job? With the BWT analysis, we can make a conclusion that carries a stronger policy implication for addressing the poor retention rates; the good levels of these three attributes are highly valued by respondents and can be said to make the nursing job attractive in an absolute sense. However, given the comparative results discussed earlier, we must also bear in mind that the true ranks of the salary levels may be higher than what is suggested by the max-diff estimates.

5 Discussion

We analyse stated preference data from two different best-worst choice experiments: a traditional DCE involving choices over nursing jobs (BWL or best-worst alternative) and a newer type of DCE involving choices over job attributes (BWT or best-worst attribute-level). Our findings suggest that the two methods elicit different preference estimates; the relative valuations of different attributes change to varying extents across the two sets of data on the same respondents. The key structural difference is in preferences over improvements in pecuniary and non-pecuniary attributes, with the BWL analysis

indicating much stronger preferences for pecuniary gains. Moreover, our results suggest that this difference is not due to the use of simple heuristics in BWL by respondents who wish to simplify the choice task.

We acknowledge that the underspecification of choice sets, together with the identification of additional utility parameters, may make BWT a particularly attractive alternative to traditional DCEs in some contexts. For instance, 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 traditional DCE may be hard to design, not least because different jobs are best described by different attributes. A BWT experiment as ours provides useful input by allowing identification of attribute-levels which make a nursing job attractive in an absolute sense.

However, in light of our findings, further accumulation of evidence on the comparability of BWT and traditional DCE methods is warranted before promoting the wider use of BWT. In particular, we recommend that future studies focus on exploring the implications of underspecified choice sets for respondents' answering strategies. Including additional information like "similar jobs (products) pay (cost) $XXX" in a BWT experiment may shed light on whether the sort of discrepancy found in our study is driven by insufficient contextual information or the reluctance to reveal preferences over monetary and non-monetary attribute-levels directly.

References

Beggs S., Cardell S. and Hausman J. (1981) "Assessing the potential demand for electric cars", Journal of Econometrics, 16, pp.1-19

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

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

Fok, D., Paap, R. and Van Dijk, B. (2011) "A rank-ordered logit model with unobserved heterogeneity in ranking capabilities", Journal of Applied Econometrics, doi: 10.1002/jae.1223

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

Greene, W. and Hensher, D. (2003) "A latent class model for discrete choice analysis: contrasts with mixed logit", Transportation Research Part B, 37(8), pp.681-698

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

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

Hess, S., Ben-Akiva, M, Gopinath, D., and Walker, J. (2011) "Advantages of latent class over mixture of logit models", mimeo, http://www.stephanehess.me.uk/papers/Hess_Ben-Akiva_Gopinath_Walker_May_2011.pdf

Hole, A. (2011) "A discrete choice model with endogenous attribute attendance", Economics Letters, 110(3), pp.203-205

Lancsar, E. and Louviere, J.J. (2008), "Estimating Individual level discrete choice models and welfare measures using best worst choice experiments and sequential best worst MNL", University of Technology, Centre for the Study of Choice (Censoc), working paper

Lloyd, A. (2003) "Threats to the estimation of benefit: are preference elicitation methods accurate?", Health Economics, 12, pp.393-402

Lusk, J. and Briggeman, B.C. (2009) "Food values", American Journal of Agricultural Economics, 91(1), pp.184-196

Lusk, J. and Natalie, P. (2009) "Consumer preferences for amount and type of fat in ground beef", Journal of Agricultural and Applied Econometrics, 41(1), pp.75-90

McFadden, D. (1973) "Conditional logit analysis of qualitative choice behaviour", In Frontiers in Econometrics, Zarembka P (ed.), Academic Press: New York, pp.105-142

McFadden, D. and Train, K. (2000) "Mixed MNL models for discrete response", Journal of Applied Econometrics, 15, pp. 447-470

Naude, M. and McCabe, R. (2005) Magnet hospital research pilot project conducted in hospitals in western australia. Contemporary Nurse, 20:38–55

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

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

Seago, J., Ash, M., Spetz, J., Coffman, J., and Grumbach, K. (2001) "Hospital registered nurse shortages: Environmental, patient, and institutional predictors", Health Services Research, 36:831–52

Shen, J. (2009) "Latent class model or mixed logit model? A comparison by transport mode choice data", Applied Economics, 41(22), pp.2915-2924

Train, K. (2008) "EM Algorithms for Nonparametric Estimation of Mixing Distributions", Journal of Choice Modelling , 1 (1), pp.40-69

Train, K. (2009) "Discrete choice methods of simulation:", Cambridge University Press

Van Ophen H., Stam S. and Van Praag B. (1999) "Multichoice logit: modelling incomplete preference rankings of classical concerts", Journal of Business and Economic Statistics, 17(1), pp.117-128

Yoo,.H. (2012) "Modelling the source of additional noise in rank-ordered data", mimeo

Table I: Nested Models in the LHROL

General Specification: Latent class heteroskedastic rank ordered logit (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, 1973)
$C \geq 2$ & $\sigma_c = 0, c = 1, \ldots, C$	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.*, 2011)

Notes: Both latent class logit and rank ordered logit can also be motivated directly from McFadden's (1973) random utility model for multinomial logit without invoking the sequential decision making process as required for heteroskedastic rank ordered logit; see derivations in Beggs *et al.* (1981) and Train (2009).

Table II: Average LMD coefficients (BWT data)

Variable	Weighted Average	Standard Error	Variable	Weighted Average	Standard Error
Supp mgt	5.383***	(0.197)	Abund park	2.291***	(0.230)
Sal 1250	5.290***	(0.225)	Private hosp	1.955***	(0.226)
Excell care	5.123***	(0.196)	FT hours	0.703***	(0.168)
Flex rost	4.474***	(0.204)	Limited park	0.570***	(0.173)
Sal 1100	4.278***	(0.217)	No rotation	0.388**	(0.187)
Encourage	4.221***	(0.199)	Sal 800	0.000	
3 rotations	3.988***	(0.219)	No encourage	−1.130***	(0.153)
Well equip	3.835***	(0.207)	Excess resp	−1.249***	(0.156)
Well staff	3.777***	(0.205)	Short staff	−1.274***	(0.155)
App resp	3.421***	(0.209)	Inflex rost	−1.275***	(0.149)
Flex hours	3.267***	(0.219)	Poor equip	−1.459***	(0.151)
Sal 950	2.468***	(0.277)	Poor care	−2.293***	(0.158)
Public hosp	2.347***	(0.174)	Unsupp mgmt	−2.551***	(0.158)

Notes: These estimates are derived from the latent class max-diff model with 7 classes. The weighted averages using class shares as weights are calculated and presented in decreasing order. The omitted attribute level is 'Salary 800.' Asymptotic standard errors are in parentheses. *** indicates that the parameter is significantly different from zero at a 1% level, ** at 5% and * at 10%.

Figure 1: Sample choice set in BWT including accept-or-not task

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 choice set in BWL with three hypothetical jobs

BWL choice screen

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

Figure 3: BWL and transformed BWT coefficients - latent class models

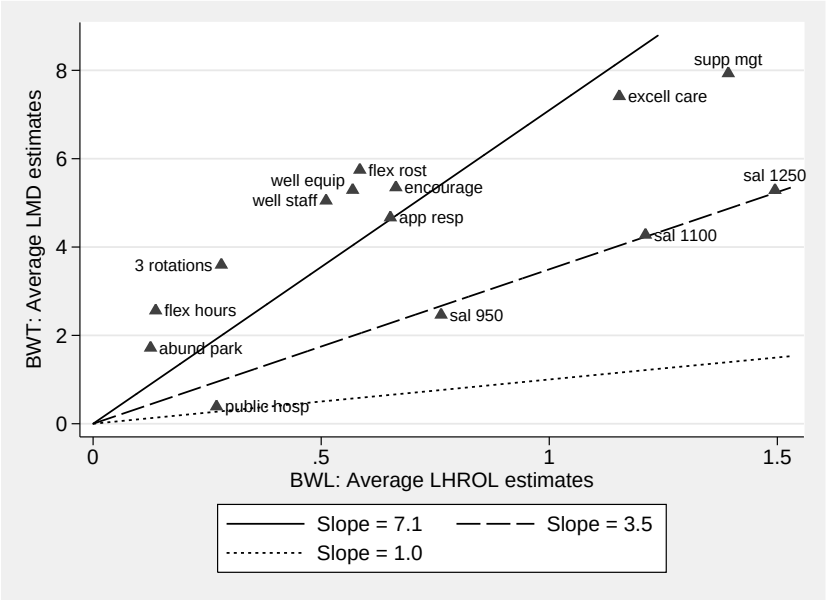


Figure 4: BWL and transformed BWT coefficients - fixed coefficients

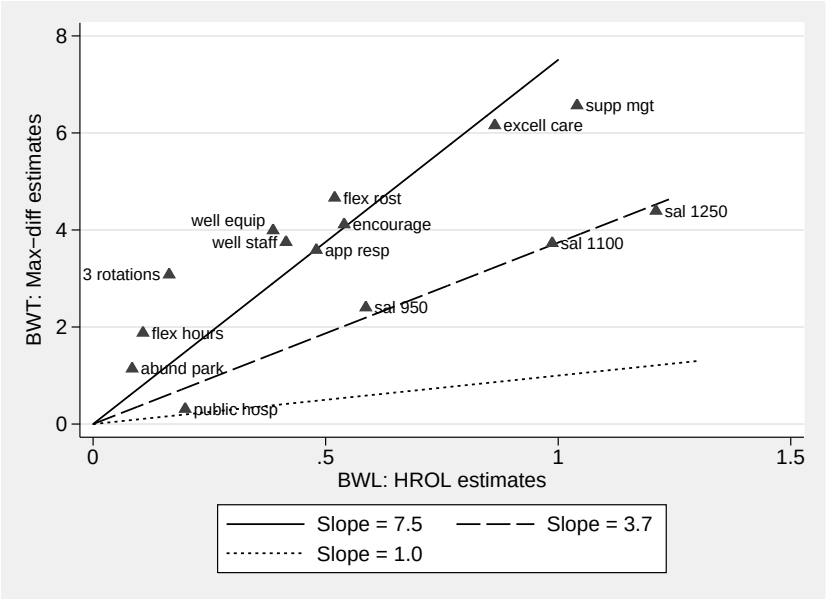


Figure 5: LHROL estimates - BWL data

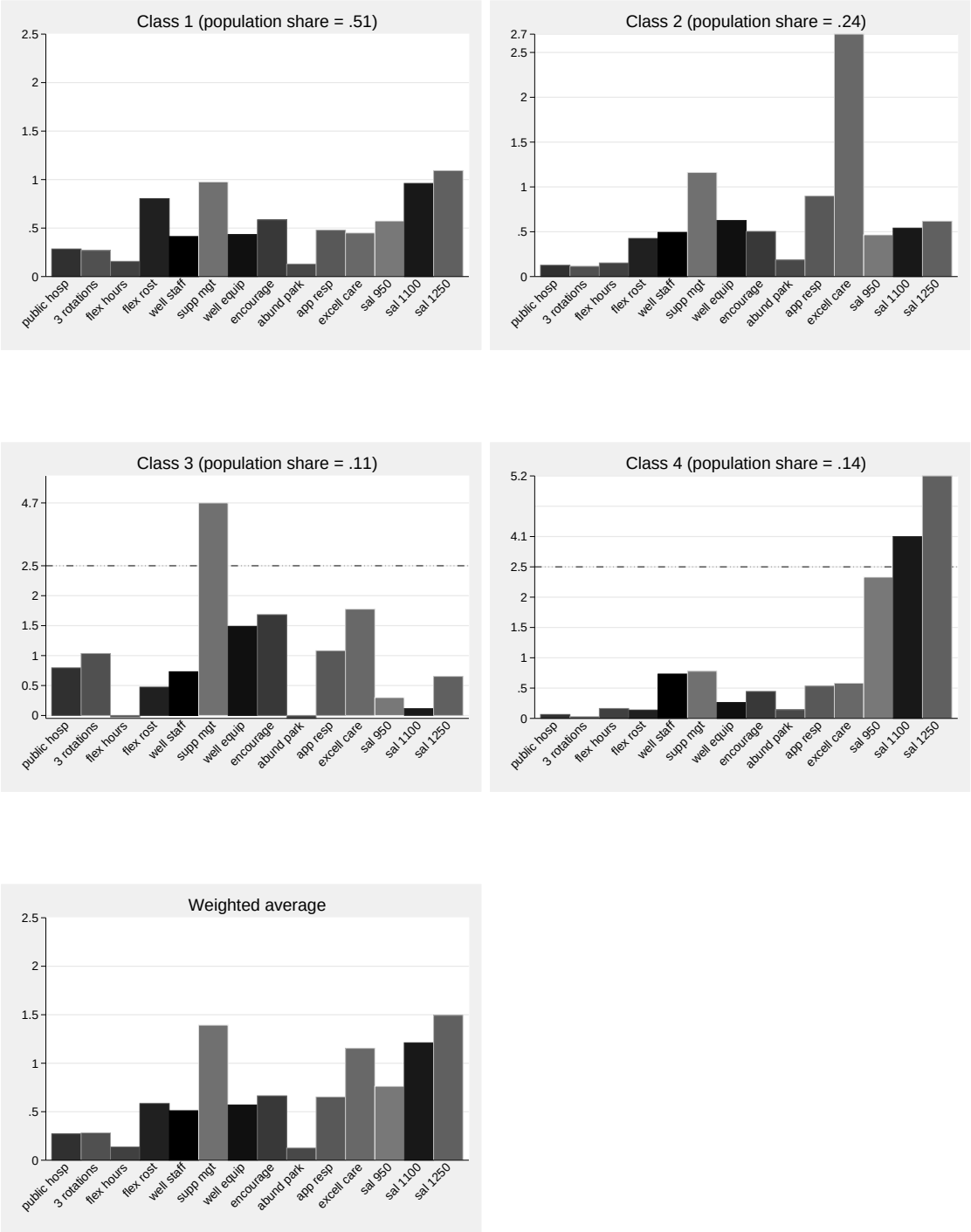


Figure 6: LMD estimates - BWT data

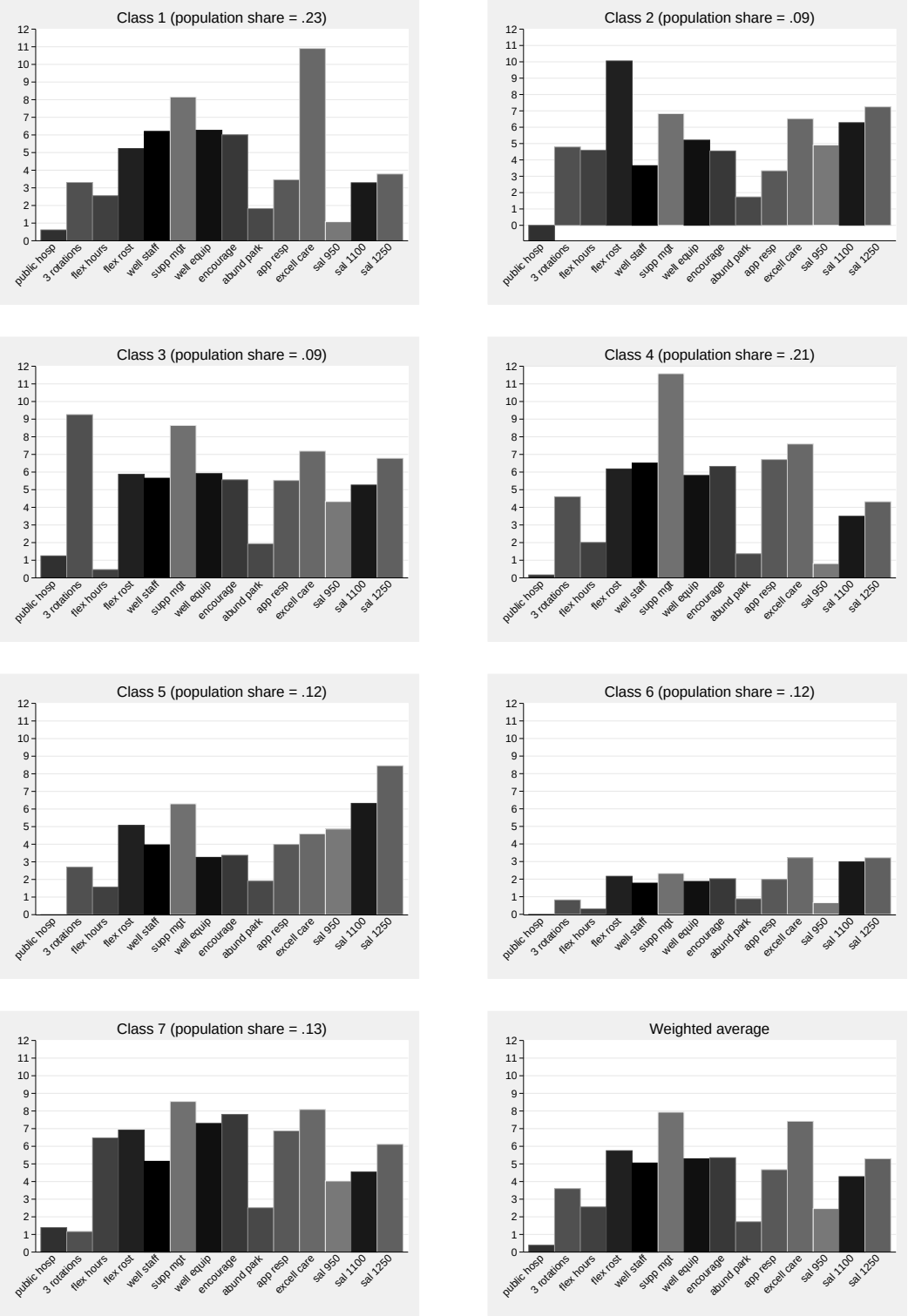


Figure 7: RE logit and latent class BWL coefficients

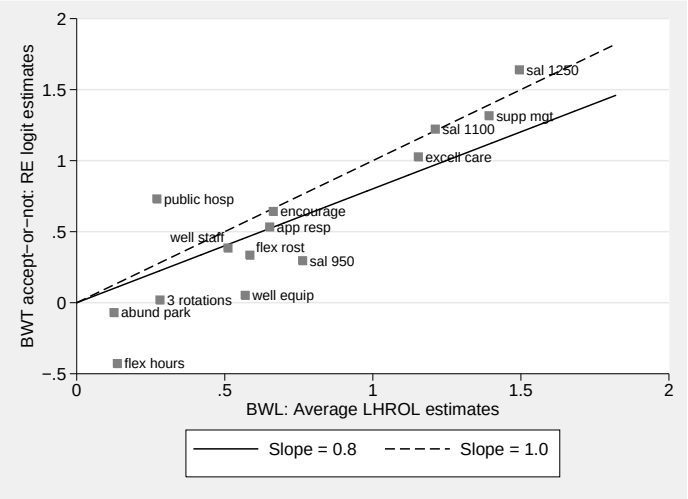
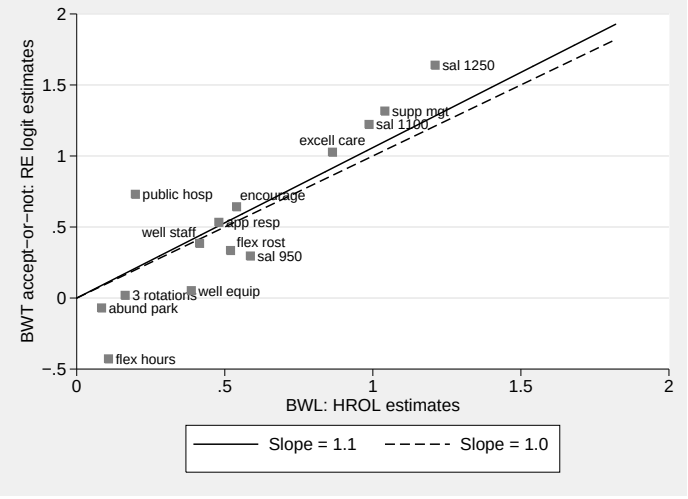


Figure 8: RE logit and fixed BWL coefficients



Appendix Table I: Attributes and Levels for the Discrete Choice Experiment and associated model variable names

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

Appendix Table II: LHROL estimation results (BWL data)

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 share	0.513***	0.241***	0.107***	0.139***	
	(0.034)	(0.031)	(0.020)	(0.019)	
Number of respondents		526	Log likelihood		-5706.48
Number of observations		21040	BIC		11832.74

Notes: The model is estimated via FIML using Stata 11.2/IC. The omitted level for salary is 800; for the other attributes, the omitted levels are provided in Appendix Table 1. σ is the ratio of the variances of the errors in the first and second steps of the ranking respectively. BIC refers to the Bayesian information criterion. Asymptotic standard errors are in parenthesis. *** indicates that the parameter is significantly different from zero at a 1% level, ** at 5% and * at 10%.

Appendix Table III: Transformed LMD estimation results (BWT data)

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 share	0.234***	0.086***	0.093***	0.210***	0.122***	0.119***	0.135***	
	(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	

Notes: The model is estimated via FIML using Stata 11.2/IC. 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. The base level for salary is 800; for the other attributes, the base levels are provided in Appendix Table 1. BIC refers to the Bayesian information criterion. Asymptotic standard errors are in parenthesis. *** indicates that the parameter is significantly different from zero at a 1% level, ** at 5% and * at 10%.