Stated preferences over job characteristics: A panel study

Denise Doiron[a]
Hong Il Yoo[b]

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

[b] Durham University Business School, Durham University, UK. h.i.yoo@durham.ac.uk

May 20, 2019

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 also grateful to Jeffrey Smith, Dwayne Benjamin, David Card and other participants of the Festschrift held in honour of Professor Craig Riddell for their comments and suggestions. We also owe special thanks to an anonymous referee and the editor for their input.

Abstract

When making choices over jobs with different characteristics, what trade-offs are decision makers willing to make? Such a question is difficult to address using typical household surveys that provide a limited amount of information on the attributes of the jobs. To address this question, a small but growing number of studies have turned to the use of stated preference experiments; but the extent to which stated choices by respondents reflect systematic trade-offs across job characteristics remains an open question. We use two popular types of experiments (profile case best-worst scaling and multi-profile case best-worst scaling) to elicit job preferences of nursing students and junior nurses in Australia. Each person participated in both types of experiments twice, within a span of about 15 months. Using a novel joint likelihood approach that links a decision maker's preferences across the two types of experiments and over time, we find that the decision makers make similar trade-offs across job characteristics in both types of experiments and in both time periods, except for the trade-off between salary and other attributes. The valuation of salary falls significantly over time relative to other job attributes for both types of experiments. Also, within each period, salary is less valued in the profile case compared to the more traditional multi-profile case.

JEL classification: C23, C25, C81, J44

Key words: preference elicitation, temporal stability, convergent validity, best-worst scaling, latent class, health workforce

1 Introduction

Stated preferences have traditionally been used to provide valuations in contexts where markets do not exist such as the valuation of environmental impacts or potential new products. Recently, non-market valuations in the form of Discrete Choice Experiments (DCE) have been used to study preferences over job characteristics in health care professions (de Bekker-Grob *et al.*, 2012). This development arose in part out of concerns over projected shortages in the health care workforce in many countries including Canada (GHWA and World Health Organization, 2014). The use of DCEs solved major data problems, namely that existing survey and registry datasets did not contain enough information on and variation in specific job attributes and choices. The information provided through DCEs allowed policy makers and employers to target certain aspects of health care employment with the goal of attracting more and better matched workers and improving retention among the current health care workforce. Examples of such studies include work on clinical workers (Kolstad, 2011), doctors (Sivey *et al.*, 2012; Holte *et al.*, 2015), midwives (Huicho *et al.*, 2012) and nurses (Yoo and Doiron, 2013). The literature is expanding, with the World Health Organization (WHO) and the World Bank now promoting the DCE approach to health workforce research (Ryan *et al.*, 2012; Araújo and Maeda, 2013). Some specialized surveys (e.g. MABEL in Australia) already include DCE components in addition to usual survey questions to gain a more comprehensive view of preferences over employment alternatives (Sivey *et al.*, 2012). We would argue that the usefulness of these methods is not restricted to health care occupations; many highly regulated jobs do not vary much in the market (e.g. education sector jobs) and more generally, standard data sources do not provide adequate information on specific job attributes other than salary.

One explanation for the limited use of DCEs and other non-market valuation methods in labor economics is the scarcity of results on the reliability of stated preferences. While there is evidence validating stated preferences using revealed preference estimates (Brownstone *et al.*, 2000; Small *et al.*, 2005), there are still many open questions regarding appropriate methods to use in various contexts and characteristics that may be specific to stated preferences. As an example, as far as we are aware, the only direct evidence on the stability of stated job preferences comes from our own studies, Yoo and Doiron (2013) and Doiron and Yoo (2017) that we will summarize shortly. In this paper, we contribute to this literature by analyzing the stability of stated preferences over job characteristics in nursing. Specifically, we study the reliability of stated prefer-

ences over time and across two elicitation methods: a traditional DCE that asks for the best and worst out of several job profiles described using particular combinations of job characteristics, and a newly proposed type of DCE known as "profile case best-worst scaling" that asks for the best and worst out of several characteristics that describe a particular job profile. This paper is the first to study the temporal stability of preferences elicited using the profile case method, and also the first to study the convergent validity of the profile case method and the traditional DCE in a longitudinal setting.

We develop a novel joint likelihood approach linking an individual's preferences across the two types of DCEs and over time, allowing the straightforward evaluation of convergent validity and temporal stability. The insight in our approach is that while the two types of DCEs present seemingly incomparable decision tasks, both types of DCEs can be modeled using generalizations of the random utility model (McFadden, 1973) where estimated coefficients correspond to preference weights over job attributes. After reparametrizations, we can compare and test restrictions on (a subset of) the preference weights in the joint model, both across DCE types and survey waves. Are the coefficients stable across methods and over time? Are they stable except for re-scaling due to latent heteroskedasticity such that the relative magnitudes of the coefficients are preserved? These are questions that can be answered in a straightforward way in our approach. In the most general specification, we model unobserved heterogeneity in preference weights assuming that each decision maker belongs to one of several possible preference "classes", and characterize each class by its own preference weights that vary freely across DCE types and survey waves. We can therefore evaluate whether there is a pattern of stability/instability in the preference weights that is observed only for a certain segment or class of decision makers. This approach could be useful to labor economists working with multinomial response data in longitudinal contexts or with data from different sources as an alternative to standard panel methods.

We now describe in more detail the DCE and the relevant literature. The DCE has rapidly increased in use since the turn of the century, overtaking contingent valuation (CV) as the most popular non-market valuation method in 2010 (Adamowicz, 2013, p.5). A traditional decision task in the DCE elicits preferences for certain attributes in a *multi-profile case* setting, by prompting choices over several profiles which represent different configurations of those attributes. For example, in the DCE that we analyze in this paper, each profile is an entry-level nursing position and attributes are relevant job characteristics such as salary, the number of clinical rotations and support for professional development. The prototypical DCE task is a multi-profile case *best* task

in which the respondent is asked to pick the most preferred option from several profiles, this choice being modeled as the utility maximizing profile. We consider a variant of this decision task that asks the respondents for their best and worst choices rather than the most preferred option only. This is known as "multi-profile case best-worst scaling (BWS)" and it has the advantage of collecting more information from a given sample size whilst exacting a minimal increase in cognitive effort.[1]

Recently, researchers in marketing and psychology proposed an alternative type of decision task that elicits preferences in a single-profile case setting, by prompting choices over attributes of a given profile. This new type of task has become increasingly popular amidst claims that it is cognitively easier for respondents and allows the identification of more preference parameters (Flynn *et al.*, 2007; Flynn, 2010), but little is known about its convergent validity and temporal stability properties. This alternative method called "profile case BWS" is a fundamentally different alternative to the prototypical task; it presents an individual with only one profile and asks her to state the best attribute and the worst attribute of that profile. To emphasize the distinction between multi-profile and profile cases in terms of information presented, Yoo and Doiron (2013) call the latter case "single-profile" and this paper uses the same label. The multi-profile and single-profile cases have been used in a range of empirical applications across disciplines, as reviewed in a recent monograph by the developers of BWS (Louviere *et al.*, 2015).[2]

The choice experiments analyzed in this paper were administered as part of a larger longitudinal survey described in Kenny *et al.* (2012), which was designed to study the job preferences of junior nurses and nursing students in Australia.[3] In each wave of the survey, every respondent was asked to complete 8 choice scenarios in the single-profile case BWS task and further 8 choice scenarios in the multi-profile case BWS task. In each single-profile case scenario, the respondent evaluated one entry-level

[1]Do people choose their most preferred options? This question is addressed in a namesake study by Caparros *et al.* (2008) and a follow-up study by Akaichi *et al.* (2013). Both studies elicit preferences by asking one group of respondents to make a choice from several profiles, and the other group of respondents to rank the same profiles from best to worst. They find that models estimated using the choice data give the same results as models estimated using the best profile data (they do not utilize information on other preference ranks), as long as the number of profiles under consideration is small say, four or less.

[2]The single-profile case BWS and the multi-profile case BWS are sometimes called case 2 BWS and case 3 BWS, respectively. As these alternative names suggest, there is also a third case of BWS known as the object case BWS or case 1 BWS. This case is very similar to the single-profile case BWS, except that attributes do not have levels that vary from profile to profile. For an empirical application of the object case in economics, see Lusk and Briggeman (2009).

[3]Yoo and Doiron (2013) and Doiron *et al.* (2014) also provide more information on the survey and the selection of attributes.

nursing job described by 12 characteristics (attribute-levels) and chose the best and worst characteristics of that job. In each multi-profile case scenario, the respondent evaluated a set of three such jobs and chose the best and worst jobs from that set. The longitudinal dimension of the survey and the within-respondent allocation of the two distinct tasks give us a unique opportunity to study the temporal stability of stated preferences by method of elicitation. Are discrepancies in preferences across method due to framing effects that persist over time or momentary behavioral noises? Addressing such questions inevitably requires a study design that has the same group of individuals completing both decision tasks at more than one point in time, and we are not aware of any other DCE which has this design.

There is a small but growing number of studies on the stability of preferences between the single-profile task and the multi-profile task, though none of them addresses temporal stability. Two previous studies elicit preferences for health states and social wellbeing, and find evidence of convergent validity (Potoglou et al., 2011; Flynn et al., 2013). We call this "intercase stability" to emphasize that the comparison is between the single-profile case task and the multi-profile case task at a point in time. More recent studies find limited evidence of intercase stability both in preferences for health states (Krucien *et al.*, 2016) and new medical technologies (Whitty *et al.*, 2014) but the findings do not point to a particular pattern of discrepancies. In contrast to most non-market valuation studies that include some measure of price or net income, these four studies only involve non-pecuniary attributes. So far, our earlier contribution (Yoo and Doiron, 2013) is the only study on intercase stability that consider the usual profile configuration involving both pecuniary and non-pecuniary attributes. In that study, we focus on the first wave of the longitudinal DCE analyzed in this paper, and find that there is an important distinction between the two types of attributes: the use of the single-profile DCE led to apparent undervaluation of salary relative to other attributes, whereas there is much greater intercase stability in the relative valuation of non-pecuniary job characteristics.

A slightly larger number of DCE studies examine temporal stability of preferences, but all of them focus on the multi-profile case task. Earlier studies consider profiles describing medical services or social care options, and find evidence of stability over relatively short horizons of four months or less (Bryan *et al.*, 2000; San Miguel *et al.*, 2002; Salkeld *et al.*, 2005; Ryan *et al.*, 2006; Skjoldborg *et al.*, 2009). Most of the recent studies originate from the environmental valuation literature, and also find evidence of stability over a short horizon (six months) (Czajkowski *et al.*, 2014; Rigby *et al.*, 2015)

but more mixed evidence when the horizon is extended to one year (Liebe *et al.*, 2012; Schaafsma *et al.*, 2014). In our previous work (Doiron and Yoo, 2017), we focus on the multi-profile case component of the nursing longitudinal survey, and compare preference weights across two waves of data spaced 15 months apart on average.[4] As in the earlier study on intercase stability, we find that the distinction between pecuniary and non-pecuniary attributes affects conclusions: over time, there was a decline in preferences for salary relative to other attributes, whereas preferences for non-pecuniary attributes remained mostly stable.

Our previous findings raise several questions for the current analysis of temporal stability by method of elicitation. As summarized above, we found that preferences for salary were not only unstable between the two cases of the DCE within the same wave (Yoo and Doiron, 2013) and between the two waves of the same task (Doiron and Yoo, 2017), but the sign of the discrepancies is intriguing. In wave 1, salary was undervalued in the single-profile case task relative to the multi-profile case task. Within the traditional multi-profile case task, salary was undervalued in wave 2 relative to wave 1. A finding that the relative undervaluation of salary in the single-profile DCE persists into wave 2 would support the presence of systematic framing effects and also a genuine decline in preferences for salary relative to other job characteristics over time. Alternatively, a finding of temporal stability in the salary weight for the single-profile DCE could result in consistent preferences across the two methods in wave 2 and suggest more instability in monetary valuations in the traditional form of DCEs. Note that in most non-market valuation studies, the primary parameter of interest is the willingness to pay (WTP) for certain attributes; WTP is the most common measure used to compare results across studies and to derive policy implications from stated preference estimates. It is computed by using the marginal utility of money to divide and monetize the utility weights on other attributes (Revelt and Train, 1998; Layton and Brown, 2000; Small *et al.*, 2005); hence, instability in the estimated weight on a monetary attribute like salary deserves special attention.

[4] As usual for an online longitudinal survey, the exact interval between the two waves varied from respondent to respondent. We note that the minimum interval was a full 12 months. We are aware of only one study (Islam and Louviere, 2015) that considered a longer horizon than ours. In that marketing application, the authors test the stability of preferences for household consumables (e.g. toothpastes) over a horizon of two years and find evidence of stability in raw best-worst responses. Unlike other studies cited above, however, they do not test the stability of the structural parameters of a random utility function.

As mentioned above, we apply an econometric framework that allows the joint analysis of DCE responses from different decision tasks and survey waves. Perhaps surprisingly, the linking of a person's preferences across decision tasks had yet to be formally recognized in the literature on intercase stability. All existing studies, including our previous work, estimated a separate model for each type of task, effectively analyzing the data as if the sample of respondents varied from task to task. The joint estimation framework developed in this paper explicitly accounts for the fact that each respondent completed both tasks in both survey waves, while also allowing for potential variations in her preferences between tasks and across waves. The workhorse models of choice behavior in the multi-profile and single-profile case tasks are the heteroskedastic rank-ordered logit model of Hausman and Ruud (1987) and the max-diff model of Marley and Louviere (2005), respectively. To operationalize our econometric framework in the context of unobserved heterogeneity, we combine these behavioral models with the non-parametric specification of unobserved heterogeneity due to Heckman and Singer (1984), and take a full set of task- and wave-specific preference parameters as a point in a discrete distribution.

Moving on to results, we find that for both single-profile and multi-profile cases, the preference weights are jointly statistically different across waves. This is true whether the test compares unrestricted models with models where all weights are forced to be equal across waves or with models where changes over time are allowed but in the scaling factor only.[5] The biggest change over time comes from the weight on salary which drops significantly across the two waves; for example, in the model without unobserved heterogeneity, the weight on salary falls by 34% in the single-profile case and 22% in the multi-profile case. There are few significant changes in the other coefficients over time and only two out of 32 parameters shift significantly at a 1% level across the waves. Excluding salary, we still reject the joint equality of preference weights across waves for both cases. Nevertheless, when we compare the relative magnitude of these parameters over time, we see a large degree of stability and in particular, we would make the same policy recommendations based on the wave 2 non-pecuniary preference weights as for

[5]In the literature, preferences have been treated as stable if they differ only through shifts in the scaling factor (the error variance). This is perhaps more compelling when comparing estimates across methods, in this case we would not expect the stochastic terms to have equal variances given the different format of the experiments and the different models used to study the stated choices. Nevertheless we also treat preferences across time as stable if estimated parameters differ only through a shift in the scaling factor; note that shifts in the scaling factor only leave the marginal rates of substitution unchanged.

the wave 1 weights. This was the conclusion for the multi-profile case in Doiron and Yoo (2017) and we find that this is also true for the single-profile case. The downward shift in preferences for salary relative to other job characteristics is substantial. For the single-profile case, salary drops from 5th to 9th place when the 12 attributes are ranked from biggest (1st) to smallest (12th) based on the sizes of the associated utility weights. For the multi-profile case, the drop in the utility weight on salary is also largest (in absolute value) among all attributes but in this case, salary remains in 2nd place in the utility weight ranking despite the drop.

When comparing results across single-profile and multi-profile cases, we find that there is a great degree of stability across the two types of DCE tasks in preferences, except for salary. Salary in the single-profile case is substantially reduced in weight relative to other attributes. This is what we had found in our earlier work (Yoo and Doiron, 2013) for the wave 1 data and the evidence here suggests that the same holds for wave 2. None of the changes across waves in the attribute weights are significantly different for the two cases at the 1% level of significance except for salary. The drop in the salary weight is more pronounced in the single-profile case; this intercase difference is the largest in magnitude across all attributes and it is significant at 1%.

In the version of the model with unobserved heterogeneity, we find that the optimal number of classes is 4 with almost equal shares across them. The latent classes can be described roughly by the relative importance each class places on the various job characteristics. One class cares more about the level of patient care, another one cares more about the other non-pecuniary job characteristics such as the management style, whether the environment is supportive and the level of responsibility is appropriate. The last two classes place more weight on salary. The main results discussed in the previous paragraphs hold for each class; our main findings are not due to one class' behavior but are found across the board.

In conclusion and as one of the authors of this paper (Doiron), I wish to acknowledge the considerable influence that Professor Craig Riddell had on my career. I was fortunate to take a course in industrial relations from Craig in my first year as a graduate student at UBC. Craig's mastery of the area and his enthusiasm for the subject converted several students to the field myself included; I ended up doing my PhD under Craig's supervision in industrial relations. The breadth of Craig's contributions both in academic and policy areas, and his thoughtful, considered views made him a terrific supervisor, mentor and later on co-author. I owe him a huge debt of gratitude. Stated preference work may seem far from typical research in the field of industrial relations

but in fact the goals of this research are closely aligned with much of Craig's work: a better understanding of workers' job preferences and well-being and a more effective design of labor contracts regulating these jobs. The fact that the work involves the nursing profession is also a testament to the tremendous impact that Rosemarie Riddell had on the betterment of the lives of the largest of the healthcare provider group and the patients they serve.

The remainder of the paper is organized as follows. Section 2 provides further information on the DCE that we analyze. Section 3 presents our econometric methods. Section 4 reports the estimated preference parameters. Section 5 discusses and concludes. The appendix section is available online at the journal website.

2 Discrete Choice Experiment

The discrete choice experiment (DCE) that we analyze was administered as part of a larger longitudinal study on the training and job decisions of junior nurses in Australia. As in many other countries, projected shortages of nursing professionals in Australia are alarming. This has led to the demand for more empirical work on the preferences over nursing job characteristics, the goal being a greater understanding of these preferences to inform more effective recruitment and retention of nurses. The underlying survey recruited 628 respondents during 2008-2010, from 3-year Bachelor of Nursing programmes at the University of Technology Sydney and the University of New England. Both institutions are large universities located in the state of New South Wales. In addition to the DCE that we will describe shortly and questions on labor market outcomes and job-related attitudes, the survey includes standard questions on demographics, socio-economic status, health and social wellbeing. A broader discussion of the policy background and the survey design are available in Kenny *et al.* (2012) and Doiron *et al.* (2014).

This paper focuses on 234 respondents who participated in the DCE involving entry-level nursing jobs in two consecutive survey waves. They completed the first-wave DCE between September 2009 and July 2011, when 27%, 32% and 41% of them were third-year, second-year and first-year students; and the second-wave DCE between April 2011 and August 2012, when 35% of them were graduate nurses, while 34%, 29% and 2% were third-year, second-year and first-year students. Each respondent's completion dates were spaced at least a full year apart, and 15 months on average.

As described above the DCE involves two distinct types of best-worst scaling decision tasks called single-profile case BWS and multi-profile case BWS. In the current context, a profile is an entry-level nursing job that is described using a certain configuration of salary and 11 other attributes. In the multi-profile case BWS task, each choice scenario asks the respondent to identify the best job and the worst job out of three distinct jobs. This task closely resembles the prototypical DCE task that would prompt the choice of the best from several profiles. The single-profile case BWS task presents much less information; each choice scenario asks the respondent to evaluate only one job, and identify the best attribute and the worst attribute of that job. Figure 1 and figure 2 show sample choice screens for the two types of BWS tasks.

[Figure 1 about here]

[Figure 2 about here]

In each wave of the underlying survey, the respondent faces 8 single-profile case choice scenarios and 8 multi-profile case choice scenarios. The DCE design and protocol remained the same between the two waves, apart from two differences. First, the set of possible salary levels for nursing jobs changed from {\$800, \$950, \$1100, \$1250} in wave 1 to {\$900, \$1100, \$1300, \$1500} in wave 2, so that the jobs looked realistic relative to the pay scales in use at the times of the launching of the two waves. Second, the respondent was required to complete all choice scenarios in wave 1, whereas the respondent was free to quit the DCE at any stage in wave 2. This quit option, however, was used by only two respondents who respectively completed 1 and 3 scenarios in the multi-profile case task before quitting; like other respondents, they chose to complete all 8 choice scenarios in the single-profile case task.

Table 1 summarizes all 12 attributes used in defining the nursing jobs, alongside the possible levels of each attribute. As summarized in Doiron *et al.* (2014), the selection of the job attributes was informed by the existing empirical literature, particularly on "magnet hospitals" in the US, and also by pilot experiments and focus group discussions. The full set of choice scenarios for inclusion in the DCE, as well as particular blocks of those choice scenarios that respondents faced, were constructed using the techniques of Street *et al.* (2005) and Street and Burgess (2007). In terms of the D-optimality criterion, the resulting design of the multi-profile case BWS task is optimal and that of the single-profile case BWS task is as good as the complete factorial design (Street

and Knox, 2012). A further summary of the DCE design process is available in Yoo and Doiron (2013).[6]

By jointly modeling the two waves of data from single-profile and multi-profile case BWS, this paper substantively extends three previous studies that analyzed the DCE component of the same survey. The focus of Doiron *et al.* (2014) was on the multi-profile case task in wave 1 and the policy implications from the resulting stated preference estimates; the study did not consider the stability of preferences between the two cases of BWS or over time. Yoo and Doiron (2013) analyzed intercase stability in wave 1, but did not consider temporal stability of either task. Moreover, in that study, we followed the literature and modeled the data as if the sample of respondents varied from task to task, without formally linking the same respondent's preferences in one type of task to another. Finally, Doiron and Yoo (2017) tested temporal stability in the context of the multi-profile case BWS, but not in the context of the single-profile case BWS. In addition, by writing down the likelihood of observing the best job without incorporating information on the worst job, the study modeled the multi-profile case BWS responses as one would model stated choice responses. The current paper presents the analysis of temporal stability by method of elicitation using all information provided in the survey (best and worst choices), which is unique both in terms of scope and econometric methodology.

3 Econometric Methods

The econometric analysis in this paper makes use of data from two distinct types of best-worst scaling (BWS) decision tasks, single-profile case BWS and multi-profile case BWS. The data span two survey waves. In each wave, each decision task provides up to 8 choice observations per respondent. It is useful to begin by focusing on the fundamentals of the models for each task. [7] Following this, we expand the model to accommodate a non-parametric specification of unobserved preference heterogeneity across individuals, as well as within-individual variations in preferences over time. We conclude the section with further comments on empirical implementation issues. Where relevant, we use the term "attribute-level" to describe a pair of an attribute and one of its possible levels. For example, consider the hospital's reputation for quality of care,

[6] We thank Debbie Street for the statistical design of all DCE tasks in the study.

[7] The task-specific models are developed in more detail in our earlier study (Yoo and Doiron, 2013) and we acknowledge some overlap in the technical discussion that follows.

which has "poor" and "excellent" levels. This attribute then generates two attribute-levels, namely poor quality of care and excellent quality of care.

3.1 Baseline Model Specifications

For the analysis of the multi-profile case BWS, the heteroskedastic rank-ordered logit (HROL) model of Hausman and Ruud (1987) is by far the most influential behavioral model. For the analysis of the single-profile case BWS, the max-diff model of Marley and Louviere (2005) has the same status. In our earlier study, we combined each of these workhorse models with the non-parametric specification of unobserved preference heterogeneity due to Heckman and Singer (1984). We also showed that the resulting latent class HROL (LC-HROL) and latent class max-diff (LMD) models nest or closely approximate many other popular models; for a broader discussion of related modeling issues, see Yoo and Doiron (2013). In that study, we estimated LC-HROL and LMD separately, effectively modeling the data as if the sample of respondents varied from task to task. In this paper, we will use LC-HROL and LMD as a basis for a joint model which recognizes that the same respondents performed both tasks.

First, we specify the HROL and max-diff models that will form the kernels of the joint likelihood function. To focus on essentials, suppose for the time being that there is only one wave of data to analyze. Let $n = 1, \ldots, N$ denote a respondent, $t = 1, \ldots, T$ a choice scenario, $k = 1, \ldots, K$ an attribute, and $l_k = 1_k, 2_k, \ldots, L_k$ a level of attribute k.[8] Each profile or job j is described by K attributes set at specific levels. Let $x_{njt}^{l_k}$ denote 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 HROL model for the multi-profile case BWS task assumes that respondent n ranks three jobs in two statistically independent steps indexed by $r \in \{f, s\}$. In the first step or step f, she picks the best of the three jobs. In the second step or step s, she eliminates her first-best job from consideration, and picks the best of the other two jobs. Note that given a choice set of three jobs, observing the respondent's best and worst jobs is equivalent to observing her best and second-best jobs, in the sense that both sets of information imply the same preference ordering. The sequence of choices assumed here is not at odds with the best-worst response format.

[8]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 in the first-wave DCE, 1_k, 2_k, 3_k and 4_k are \$800, \$950, \$1,100 and \$1,250 respectively.

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 stochastic behavioral error term. Specifically, for $r \in \{f, s\}$

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

where u^f_{njt} and u^s_{njt} are independently extreme value distributed with variances equal to $\pi^2/6$ and $\pi^2/(\sigma^2_n 6)$ respectively.[9] $B^{l_k}_n$ measures person n's utility of having attribute-level l_k and its scale has been implicitly normalized along with the variance of u^f_{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 normalized to 0, giving identified parameters $\beta^{l_k}_n = B^{l_k}_n - B^{1_k}_n$ for $l_k = 2_k, \ldots, L_k$. 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 response in scenario t of the multi-profile case BWS. Given the stochastic assumptions, 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 stated that the best job is job 1 and the worst job is job 3, the likelihood becomes

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

[9]Since $u^f_{njt} \neq u^s_{njt}$, in the HROL framework, the respondent uses one set of utility functions to identify the best job and another set of utility functions to identify the second best job from the same choice set t. From a microeconomic perspective, it may be more natural to write out a model that has the respondent use one set of utility functions, say $U_{njt} = \boldsymbol{\beta}_n \cdot \mathbf{x}_{njt} + \varepsilon_{njt}$ where the error term ε_{njt} is *i.i.d.* extreme value, to rank all alternatives in choice set t from best to worst. This is the random utility model that Beggs *et al.* (1981) have specified in their seminal study to develop the rank-ordered logit (ROL) model. Interestingly, even though ROL does not assume two-step decision making as HROL does, the conditional independence properties of the extreme value distribution implies that the ROL probability is a special case of the HROL probability in equation (2) that occurs when $\sigma_n = 1$. The popularity of HROL over ROL stems from that empirical studies typically reject the hypothesis of $\sigma_n = 1$. As a matter of fact, Hauman and Ruud (1987) have proposed HROL specifically to address empirical regularities that in ROL applications, the estimates of $\boldsymbol{\beta}_n$ tend to become more attenuated when one uses data on deeper preference ranks; that is, the estimates become smaller in magnitude when one uses data on first and second preferences instead of first preferences only, when one uses data on first, second and third preferences instead of first and second preferences only, and so on.

where σ_n measures heteroskedasticity across steps in the ranking. The likelihood of other responses takes the same functional form, with obvious permutations of job indices. The heteroskedasticity parameter captures the notion that people may 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 behavioral errors (Hausman and Ruud, 1987). If there is more preference uncertainty in the second step response, σ_n will lie in the $(0, 1)$ interval. Special cases include preferences such that person n ranks all jobs equally systematically ($\sigma_n = 1$) or picks the second-best job arbitrarily ($\sigma_n = 0$).

The max-diff model for the single-profile case BWS task assumes that respondent n explodes K attribute-levels that make up a profile into $K(K-1)$ pairs of best and worst attribute-levels. The functional form of the max-diff model is equivalent to a multinomial logit model that regards the respondent's stated best-worst pair as the most preferred out of such $K(K-1)$ pairs.

More specifically, assume that respondent n evaluates each candidate pair by considering the utility difference between the component best and worst attribute-levels, and chooses a pair that maximizes such difference (hence "max-diff").[10] To formalize the idea, let $A_n^{l_k}$ denote respondent n's systematic utility from attribute-level l_k.[11] Each utility difference can be decomposed into systematic and stochastic behavioral error components. In case the candidate pair of interest is one that states that attribute q is

[10] One may argue that the max-diff model is descriptively implausible when the number of attributes is large: for example, with 12 attributes as in the current context, a max-diff respondent would consider 132 ($= 12 \times 11$) best-worst pairs in each scenario. Alongside the max-diff model, Marley and Louviere (2005) introduce a sequential best-worst model that may be considered more descriptively plausible: it assumes that the respondent initially chooses the best out of 12 attributes and then proceeds to choose the worst out the remaining 11 attributes. From the empirical practitioner's perspective, however, there are very limited reasons to prefer one model to the other, and we opt for the max-diff model that is the workhorse model in the literature; as we point out in Yoo and Doiron (2012, p.18), the two behavioral models lead to algebraically similar likelihood functions and consequentially similar empirical results.

[11] We change the notation for utility weights from $B_n^{l_k}$ to $A_n^{l_k}$ to emphasize that their scale is normalized with respect to potentially different error 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.

the best and attribute h is the worst, the resulting utility difference $D_{nt}^{\{q,h\}}$ is

$$\begin{aligned} D_{nt}^{\{q,h\}} &= \sum_{l_q=1_q}^{L_q} \sum_{l_h=1_h}^{L_h} (A_n^{l_q} - A_n^{l_h}) x_{nt}^{l_q} x_{nt}^{l_h} + e_{nt}^{\{q,h\}} \\ &= \sum_{l_q=1_q}^{L_q} \sum_{l_h=1_h}^{L_h} (\alpha_n^{l_q} - \alpha_n^{l_h}) x_{nt}^{l_q} x_{nt}^{l_h} + e_{nt}^{\{q,h\}} \end{aligned} \tag{3}$$

where the error term $e_{nt}^{\{q,h\}}$ is independently type I extreme value distributed. We omit profile subscript j from attribute-level dummies $x_{njt}^{l_k}$ since each scenario involves only one profile. To achieve identification, one utility parameter needs to be normalized to 0, say for the first level of the first attribute $A_n^{1_1}$. Then, each identified parameter can be defined as $\alpha_n^{l_k} = A_n^{l_k} - A_n^{1_1}$, and now $\alpha_n^{l_k} > \alpha_n^{l_l}$ for two different attributes k and l implies $A_n^{l_k} > A_n^{l_l}$. In this sense, the single-profile case BWS allows one to learn more about the underlying preferences than the multi-profile case BWS.

Let $F_{nt}(\boldsymbol{\alpha}_n)$ denote the likelihood of respondent n's stated best-worst pair in scenario t of the single-profile case BWS task where $\boldsymbol{\alpha}_n$ is the vector of identified parameters. Given the stochastic assumptions, once the identified parameters $\boldsymbol{\alpha}_n$ are known, this likelihood takes the multinomial logit form where the index of each alternative equals the systematic utility difference within a candidate best-worst pair. For instance, in case respondent n has stated that the best attribute is q and the worst attribute is h, this likelihood becomes

$$F_{nt}(\boldsymbol{\alpha}_n) = \frac{\exp(\sum_{l_q=1_q}^{L_q} \sum_{l_h=1_h}^{L_h} (\alpha_n^{l_q} - \alpha_n^{l_h}) x_{nt}^{l_q} x_{nt}^{l_h})}{[\sum_{k=1}^{K} \sum_{l \neq k} \exp(\sum_{l_k=1_k}^{L_k} \sum_{l_l=1_l}^{L_l} (\alpha_n^{l_k} - \alpha_n^{l_l}) x_{nt}^{l_k} x_{nt}^{l_l})]} \tag{4}$$

where $l \neq k$ means that the summation is over $l \in \{1, 2, \cdots, K\} \setminus \{k\}$.

3.2 Temporal Variation and Preference Heterogeneity

To accommodate between-wave variations in preferences and data, more notation is needed. We use $w = 1, 2$ to index a survey wave, and assume that within each type of BWS task, the respondent faced scenarios $t = 1, \ldots, 8$ in wave 1 and scenarios $t = 9, \ldots, 16$ in wave 2.[12] In the remainder of this section, the discussion will proceed as if

[12]Strictly speaking, the range of the choice scenario index should have been modified to address that 2 out of 234 respondents did not complete all 8 multi-profile case scenarios in wave 2. We overlook this minor technical inaccuracy to avoid introducing extra notation.

all terms in equation (1) through equation (4) had been indexed by wave subscript w; for example, in the multi-profile case BWS, the vector of identified preference parameters in wave w is $\boldsymbol{\beta}_{nw}$, with a typical element β_{nw}^{lk}. We will abuse notation somewhat by using $P_{nt}(\boldsymbol{\beta}_{nw}, \sigma_{nw})$ to denote the HROL likelihood of any actually stated response, although equation (2) refers to a particular ranking of three jobs. Likewise, we use $F_{nt}(\boldsymbol{\alpha}_{nw})$ to denote the max-diff likelihood of any actually stated response, although equation (4) refers to a particular best-worst pair.

Conditional on individual-specific preference parameters, the joint likelihood of observing an entire sequence of 16 single-profile case responses and 16 multi-profile case responses by respondent n over two waves can be written as

$$Q_n(\boldsymbol{\theta}_n) = \left[\prod_{t=1}^{8} P_{nt}(\boldsymbol{\beta}_{n1}, \sigma_{n1}) \times \prod_{t=1}^{8} F_{nt}(\boldsymbol{\alpha}_{n1})\right] \times \left[\prod_{t=9}^{16} P_{nt}(\boldsymbol{\beta}_{n2}, \sigma_{n2}) \times \prod_{t=9}^{16} F_{nt}(\boldsymbol{\alpha}_{n2})\right] \quad (5)$$

where $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_{n1}, \sigma_{n1}, \boldsymbol{\beta}_{n2}, \sigma_{n2}, \boldsymbol{\alpha}_{n1}, \boldsymbol{\alpha}_{n2})$ is the vector of all identified parameters across decision tasks and waves.

In the first section of the results below, we present estimates of a baseline model where preferences are assumed to be homogeneous across respondents.[13] This is followed with results from a framework where preference heterogeneity across respondents is taken into account. With unobserved heterogeneity, the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}_n$ is consistent when the number of choice scenarios per each type of BWS task goes to infinity in every wave, a condition that is at odds with the current data environment. To estimate models of unobserved interpersonal heterogeneity consistently using short panel data such as ours, the researcher may specify the individual-specific preference parameters as draws from a population distribution and estimate that distribution. We adopt this random parameter modeling approach.

We apply the technique of Heckman and Singer (1984) by using a discrete distribution with C support points to approximate population heterogeneity in preferences non-parametrically. Put another way, we assume that the population consists of C types or "classes" of individuals, where each type c has its own preference vector $\boldsymbol{\theta}_c$ $=(\boldsymbol{\beta}_{c1},\ \sigma_{c1},\ \boldsymbol{\beta}_{c2},\ \sigma_{c2},\ \boldsymbol{\alpha}_{c1},\ \boldsymbol{\alpha}_{c2})$. Let $\eta_c = \Pr(\boldsymbol{\theta}_n = \boldsymbol{\theta}_c)$ denote the population share of class c, which corresponds to the relative frequency of class c in the respondent population. The unconditional joint likelihood of respondent n's responses can be then

[13]Models with homogeneous preferences are estimated with Stata commands clogit and clogithet. The latter is a user-written Stata command by Arne Risa Hole (Hole, 2006).

specified by taking the expected value of (5) as follows

$$L_n(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C; \eta_1, \ldots, \eta_{C-1}) = \sum_{c=1}^{C} \left[\eta_c \times Q_n(\boldsymbol{\theta}_c)\right] \tag{6}$$

where η_C is omitted from the argument list since it is implied by the other class shares, $\eta_C = 1 - \sum_{c=1}^{C-1} \eta_c$. The MLE of parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C; \eta_1, \ldots, \eta_{C-1}\}$ that maximizes the sample log-likelihood function $\sum_{n=1}^{N} \ln L_n(.)$ is consistent in short panels, and can be conveniently computed using the EM algorithms of Bhat (1997).[14] The joint likelihood in equation (6) simplifies to the LC-HROL likelihood when all single-profile case responses are omitted from the data set and all identified parameters are constrained to be identical between waves. The LMD likelihood can be similarly obtained by omitting all multi-profile case responses and imposing temporal stability of all identified parameters.

This joint model allows the researcher to test a wide range of hypotheses concerning the stability of preferences over time and between the two cases of BWS. For example, consider first the temporal stability of preferences elicited using the multi-profile case BWS. In the presence of preference heterogeneity, one may approach this question by taking the expected value of $\boldsymbol{\beta}_{cw}$ over the whole discrete distribution, $\mathrm{E}(\boldsymbol{\beta}_{cw}) = \sum_{c=1}^{C} \left[\eta_c \times \boldsymbol{\beta}_{cw}\right]$, and then testing the equality of the resulting population mean parameters between waves. Alternatively, one may address the question of whether preferences tend to be stable only for certain segments of the population, by testing the equality of $\boldsymbol{\beta}_{c1}$ and $\boldsymbol{\beta}_{c2}$ on a class-by-class basis. Regardless of which approach one takes, the test can be carried out individually on the utility weight on a particular attribute-level, as well as jointly on several utility weights. An analogous procedure can be applied in the context of the single-profile case BWS, based on preference parameters $\boldsymbol{\alpha}_{cw}$. It is also straightforward to test the temporal stability of preferences elicited using both cases of BWS jointly, for instance by formulating $H_0 : \mathrm{E}(\boldsymbol{\beta}_{c1}) = \mathrm{E}(\boldsymbol{\beta}_{c1}), \mathrm{E}(\boldsymbol{\alpha}_{c1}) = \mathrm{E}(\boldsymbol{\alpha}_{c2})$.

Next, consider the stability of preferences between the two cases of BWS. To make preference parameters comparable across methods, define a parametric transformation for the single-profile case BWS, $\Delta\alpha_{cw}^{l_k} = \alpha_{cw}^{l_k} - \alpha_{cw}^{1_k}$, which captures the systematic utility difference between the l^{th} level and the first level of attribute k, just like parameter $\beta_{cw}^{l_k}$

[14] All estimation results for models with preference heterogeneity have been obtained in Stata 14.2/SE using self-written Mata programs. The relevant codes are available upon request.

of the multi-profile case BWS. Tests can be conducted on these normalized preference weights. When preferences remain stable between the two cases so that the only latent source of differences in choice behavior is the scale of the behavioral error term, the identified parameter ratio $\Delta\alpha_{cw}^{l_k}/\beta_c^{l_k}$ simply equals the ratio of the two error scales. This property implies a set of parametric restrictions (namely, the same value of $\Delta\alpha_{cw}^{l_k}/\beta_c^{l_k}$ for any attribute-level l_k) that can be tested within each class to study whether people apply the same criteria to evaluate job attributes in both cases of BWS. Such simplicity of statistical testing is a major benefit of joint estimation: if one estimates a separate latent class model for each type of experiment, one cannot formally link a class in one model to a class in another model and it is therefore not possible to test intercase stability within latent classes.

3.3 Empirical Implementation

As noted above the multi-profile case (or HROL) component of our joint model is identified by normalizing the utility weight on one level of *each* attribute to zero. Except for hospital type and clinical rotations, all attributes in our analysis are vertical attributes for which certain levels are intrinsically better than others, and we choose the worst level of each attribute for normalization. As usual, such normalization can be operationalized by omitting the relevant attribute-levels from the list of regressors. Each identified utility weight in this component then measures extra utility that a level of some attribute offers over the worst level of the same attribute. For hospital type and clinical rotations, we choose "private hospital" and "no rotation" as the omitted levels; our previous studies (Yoo and Doiron, 2013; Doiron et al., 2014; Doiron and Yoo, 2017) found that these levels gave less utility than "public hospital" and "three rotations".

The single-profile case (or max-diff) component of our model is identified by normalizing the utility weight on one level of just *one* attribute to zero. We choose "short staff", meaning that the hospital is frequently short of staff, as the omitted attribute-level so that each identified utility weight in this component measures the extra utility a certain attribute-level offers over "short staff".[15]

All discussion so far assumes that the random utility functions are specified solely in terms of attribute-level dummies. In most non-market valuation studies, however, a pecuniary attribute like salary is modeled as a continuous variable when specifying

[15]Both in this case and the multi-profile case, our estimation does not constrain any extra utility to take a particular sign.

a random utility function, rather than exploded into attribute-level dummies. The continuous variable approach allows the researcher to derive the marginal utility of money and use it to divide other utility weights, thereby monetizing them into the willingness to pay for any attribute (Revelt and Train, 1998; Layton and Brown, 2000; Small *et al.*, 2005). For this reason, our previous studies based on the first-wave of the multi-profile case BWS (Doiron *et al.*, 2014; Doiron and Yoo, 2017) specified the random utility function in terms of $\ln(salary)$ to estimate the marginal utility of money, instead of including dummies for three salary levels ($950, $1000 and $1250) to estimate the discrete utility difference between each level and the base level of $800.[16]

Following common practice, the main analysis in this paper specifies salary as a continuous variable. In the HROL component of our model, the extra utilities that the three highest salary levels offer over the base level are identified in each wave, and we replace the three salary-level dummies with $\ln(salary)$ in each wave. The utility weight on $\ln(salary)$ is allowed to change between waves, like the utility weights on other non-pecuniary attribute-levels. In the max-diff component of our model, the extra utilities that all four salary levels, including the base level, offer over "short staffing" are identified in each wave. To avoid imposing more constraints than what the $\ln(salary)$ specification implies in the context of the HROL component, in each wave we continue to include the base salary-level dummy alongside $\ln(salary)$ that replaces the three other salary-level dummies. The utility weights on both the base level dummy and $\ln(salary)$ are allowed to vary between waves.

The main analysis below focuses on the fully flexible specification that allows all utility weights to vary over time and between the two cases of BWS. This makes it easy to see which parameters are shifting significantly over time and across decision tasks, but proliferate the number of free parameters to estimate. Hensher *et al.* (1999) propose a more parsimonious approach to modeling parametric shifts. In this restricted model, it is assumed that preferences for attribute-levels remain stable over time (across decision tasks) but the variance of behavioral errors shifts between the data sources. Specifically, since identified parameters in discrete choice models are inversely propor-

[16]Technically speaking, equation (1) and our earlier discussion do not actually rule out the $\ln(salary)$ specification; it can be viewed a special case of the dummies specification that places two constraints to impose a constant utility change per each logarithmic unit increase in salary from the base level. Specifically, let $\beta_n^{\$950}$, $\beta_n^{\$1100}$, and $\beta_n^{\$1250}$ denote identified utility weights on the superscripted salary levels. The $\ln(salary)$ specification implies two restrictions: $\beta_n^{\$1100} = \beta_n^{\$950} \times [\ln(1100) - \ln(800)]/[\ln(950) - \ln(800)]$ and $\beta_n^{\$1250} = \beta_n^{\$950} \times [\ln(1250) - \ln(800)]/[\ln(950) - \ln(800)]$. Also, we note that the log-linear specification performed better than the use of a linear function for salary.

tional to the latent error variance, this shift in the latent error variance will induce proportionate shifts in all identified parameters. The researcher can impose this type of parametric stability by allowing and estimating a common factor that scales the same set of identified parameters up or down across different data sources. In fact, this approach is already being used in the HROL component of our model, where we specify the scalar factor σ_{cw} to account for behavioral differences between the best choice and the second-best choice in the multi-profile case BWS, without estimating a different set of utility weights for the second-best choice. The same modeling device may be applied more broadly to account for behavioral differences between the two cases of BWS, as well as between the two survey waves, while maintaining the stability of preference weights over attribute levels.

We also present the results of estimations involving two restricted specifications, to explore how well this parsimonious approach with stable preferences would have explained the observed choices relative to our general approach. The first specification is a stable preference model building on Hensher *et al.* (1999), that takes the utility weights for the single-profile case in wave 1 as primitives and constrains the utility weights in other data sources to be the scalar multiples of these primitives. Based on the earlier notation, this model entails the estimation of scalar factors γ_{c1}, λ_c^{SP} and λ_c^{MP} to impose the following restrictions

$$\begin{aligned} \beta_{c1}^{l_k} &= \gamma_{c1} \times (\alpha_{c1}^{l_k} - \alpha_{c1}^{1_k}) \\ \alpha_{c2}^{l_k} &= \lambda_c^{SP} \times \alpha_{c1}^{l_k} \\ \beta_{c2}^{l_k} &= \lambda_c^{MP} \times \beta_{c1}^{l_k} \end{aligned} \tag{7}$$

for each class c and attribute-level l_k, instead of estimating the left-hand side coefficients directly. Much like the parameter σ_{cw} in the HROL model, $\gamma_{c1} < 1$ indicates that the latent error variance is larger in the multi-profile case than in the single-profile case for wave 1. Other scalar factors can be interpreted correspondingly. The second restricted model is a hybrid model, which continues to impose such constraints on the coefficients on all non-pecuniary attributes but allows the coefficients on salary to vary freely as in our flexible specification. This hybrid model captures the key qualitative conclusion from our main analysis that salary is the primary domain of preference instability,

something which our earlier studies also find albeit in more limited contexts (Yoo and Doiron, 2013; Doiron and Yoo, 2017).[17]

4 Results

Our analysis is based on the joint estimation of preferences elicited using two distinct methods (single-profile case BWS and multi-profile case BWS) in two survey waves. We study three major aspects of the reliability of stated job preferences. The first aspect is the temporal stability of preferences elicited using each method, that is whether preferences remain stable between survey waves within each type of experiment. The second aspect is the intercase stability of preferences, that is whether preferences elicited using the two different methods are comparable within the same survey wave. The third aspect is the relative stability of preferences elicited using different methods, that is whether one method yields more temporally stable preferences than the other and also whether there is a distinctive pattern of intercase discrepancies that persists between waves.

When combined with the Heckman-Singer technique for approximating unobserved preference heterogeneity non-parametrically (Heckman and Singer, 1984), our econometric strategy estimates 291 parameters from a sample of 7,476 choice observations on 234 respondents.[18] Prior to presenting the results from this more general model, we discuss a more parsimonious specification that assumes homogeneous preferences across individuals. While the assumption of preference homogeneity may be overly restrictive, the simpler specification conveys a useful overview of our findings in that the representative agent identified in the simpler model shows similar patterns of preference stability as each of the preference segments identified in the preferred specification.

[17] This latter specification is closely related to joint models for stated preference and revealed preference data (Brownstone *et al.*, 2000; Small *et al.*, 2005). This type of model often assumes that utility weights on certain attributes (e.g. major product attributes) are stable between the two sources of data subject to an error scale differential, but also include other utility weights that are either identified only in one source of data (e.g. attributes omitted from the DCE for the parsimony of the experimental design) or allowed to vary between the data sources. For a primer on this empirical strategy, see Train (2009, pp. 152-156).

[18] In each survey wave, each respondent completed up to 8 single-profile case choice scenarios and 8 multi-profile case choice scenarios. In total, our sample comprises 3,744 single-profile case choice observations and 3,732 multi-profile case choice observations. The two cases have different sizes because two respondents withdrew after completing 1 and 3 multi-profile case choice scenarios in the second wave.

4.1 Homogeneous preference model

We estimate a homogeneous preference model which constrains the latent class model in equation (6) to allow for only one preference segment, by setting $C = 1$. Since the single-profile case BWS prompts choices over job attributes whereas the multi-profile case BWS prompts choices over jobs, one cannot compare observed choices directly across the two cases of BWS. The random utility framework opens the door to intercase comparisons by explaining the two sets of observed choices in terms of a common construct, namely the decision maker's latent preferences for job attributes that enter the model as utility coefficients.

We focus on coefficients measuring the extra utility that a level of one attribute offers relative to the worst level of the same attribute, as these coefficients are identified in both cases of BWS. In terms of the notation introduced in Section 3.2, they correspond to $\Delta\alpha_{cw}^{l_k} = \alpha_{cw}^{l_k} - \alpha_{cw}^{1_k}$ for the single-profile case and $\beta_{cw}^{l_k}$ for the multi-profile case, though the assumption of preference homogeneity makes the c subscript redundant. To make the coefficient on $\ln(salary)$ comparable to other coefficients that measure discrete utility differences, we derive and report the implied extra utility from a 60% increase in salary; as discussed in Section 2, the highest salary level in each survey wave was about this much larger than the lowest salary level.

Despite the growing literature on the single-profile case BWS, the temporal stability of utility coefficients estimated using this method has not been evaluated to date. The top panel of Figure 3 reports our findings on this issue, by plotting the coefficients for wave 2 ($\Delta\alpha_{c2}^{l_k}$) against the coefficients for wave 1 ($\Delta\alpha_{c1}^{l_k}$). The corresponding results for the multi-profile case are reported in the bottom panel that plots $\beta_{c2}^{l_k}$ against $\beta_{c1}^{l_k}$. The top panel is more populated than the bottom panel, as it also reports the coefficients that are only identified in the single-profile case; each of such coefficients ($\alpha_{cw}^{1_k}$) measures the extra utility that the worst level of a particular attribute offers relative to "short staff", i.e. the worst level of attribute *staffing levels*.

Figure 3 shows a remarkable degree of temporal stability in preferences elicited using the single-profile case BWS, with the exception of preferences for salary. In the absence of any between-wave change, the utility coefficients will line up along the dotted 45-degree line. In the top panel, we observe that the 21 non-salary coefficient estimates are tightly scattered around the 45-degree line, including the four estimates that show statistically significant changes (well staff, three rot, poor equip, and poor qual). Indeed, the line of best fit through these 21 estimates is practically indistinguishable from the

45-degree line, and has an R^2 value of 0.98, an intercept of 0.162 and a slope of 1.03. The estimated coefficient on salary shows a significant difference and is located relatively far below the 45-degree line, suggesting the relative undervaluation of salary in wave 2. The temporal stability of preferences in the multi-profile case is only slightly less remarkable. With the slope of 0.754 and the intercept of 0.098, the trendline fitted through the non-salary coefficients in the bottom panel of Figure 3 deviates more from the 45-degree line but still has an impressive R^2 of 0.92.

[Figure 3 about here]

Table 2 reports the underlying estimation results for Figure 3; these generally reinforce the qualitative insights provided by the graphical comparisons. For now, we focus on the first four columns of this table that report the estimates for each type of DCE task in wave 1 alongside the between-wave changes in those estimates. When it comes to the single-profile case coefficients $\Delta\alpha_{c1}^{l_k}$ and $\Delta\alpha_{c2}^{l_k} - \Delta\alpha_{c1}^{l_k}$ in Panel A of the table, we find that only a few of the non-salary coefficients shift significantly across the two waves and none of the shifts are significant at the 1% level. Although a Wald test shows that jointly, these shifts are significant (the χ^2 value is 28.49 with 11 degrees of freedom and a corresponding p-value of 0.003), they are relatively small and any policy recommendations based on the strength of preferences over job characteristics would remain in the two waves. The same cannot be said of the weight on salary. We find a significant shift down in the relative weight on salary both in terms of its magnitude and its statistical significance. The preference weight on a 60% rise in salary drops by 34% across waves and its rank in the attributes falls from 5th to 9th place.

[Table 2 about here]

The temporal stability of preference estimates in the traditional multi-profile case has been discussed more extensively in our previous work (Doiron and Yoo, 2017). The results presented in Table 2 are consistent with our previous conclusions that although preferences exhibit a high degree of stability across waves, salary stands out with the largest drop in its weight over time.[19] Finally we mention that there is evidence of heteroskedasticity across the choice of best and second best job profiles; the systematic

[19]Note that the multi-profile case columns of Table 2 deviate slightly from the corresponding table in our previous work for two reasons. First, our previous analysis did not incorporate information on

component of utility needs to be scaled down significantly ($\sigma = 0.597$ is significantly different from one at the 1% level) to explain the second best choices, suggesting that the decision makers are more certain about their first-best choices. Interestingly, there is no evidence of a shift in this scaling factor over time in any of our specifications.

Between the two waves, 35% of our respondents completed their Bachelor of Nursing degrees and started working as graduate nurses. This invites the questions of whether the drop in preference for salary in wave 2 is associated with their school-to-work transition, and more generally whether there are systematic preference changes during such transition that can be factored into effective workforce retention policies. In our earlier work (Doiron and Yoo, 2017), we explored these issues using model specifications that incorporated observed preference heterogeneity into utility weights.[20] We found no evidence that the utility weights on salary and other attributes varied systematically between graduate nurses and students, and across students in different years of study, with one exception; first-year students were more sensitive to the quality of care provided in the hospital. This lack of variance may reflect a good understanding of actual nursing jobs by students acquired through the practicum component of the Bachelor of Nursing program, which lessens the possible influence of career progression as an external shock to preferences. The hours spent in practicum placements during the 3-year programme are substantial: 120 in year 1, 320 in year 2 and 400 in year 3. Our findings lend support to the use of prospective workforce cohorts in DCE research to aid forward-looking policy formulation. For instance, Blauuw *et al.* (2010), Kolstad (2011), and Holte *et al.* (2015) administer DCEs involving entry-level positions to students of professional degree programmes, and Sivey *et al.* (2012) and Pedersen and Gyrd-Hansen (2014) administer DCEs involving specialist positions to non-specialist doctors.

We now turn to the question of intercase stability. The right-most columns of panel A in Table 2 report intercase coefficient differences in wave 1, as well as shifts in the intercase differences between waves. None of the between-wave shifts in coefficients are significant at the 1% level. Only two shifts are significantly different from 0 at the 5% level and these relate to coefficients that are in the bottom half of the list of the weights ranked by their relative magnitude. Salary is the exception, the decrease in its utility

the worst jobs. Second, Table 2 reports the results from a homogeneous preference model, whereas our previous work reported the population average results from a heterogeneous preference model.

[20]For instance, consider the utility weight on $ln(salary)$. Allowing this weight to vary with a personal characteristic can be achieved in the same manner as allowing the marginal effect of one regressor to vary with another regressor in a linear regression model. That is, our current model specification can be extended by including an interaction term between $ln(salary)$ and that personal characteristic.

weight across waves is much more important in the single-profile method and this shift in the intercase gap is significant at the 1% level.

Given the different behavioral models for the two decision tasks and the likely variations in the sources and nature of latent errors, some intercase variations in the overall scale of the estimates may be expected.[21] Since identified coefficients in discrete choice models are inversely proportional to the latent error variances, an intercase difference in the latent error variance could induce the overall scale to differ even when scale-free measures of preferences for attribute-levels remain stable. Based on this argument, the focus of intercase comparisons should be on whether the relative magnitudes of different coefficients remain stable across methods, instead of whether each individual coefficient remains stable across methods.

With this caveat in mind, we consider the top panel of Figure 4 that plots the single-profile case coefficients against the corresponding multi-profile case coefficients for wave 1. All single-profile case coefficients are larger in magnitude than the corresponding multi-profile case coefficients, suggesting that the latent error variance is smaller in the single-profile case. But when it comes to the relative magnitudes of non-salary coefficients, there is a good deal of agreement between the two methods and indeed the trendline across this panel has an R^2 of 0.85. In particular, when one coefficient is located to the right of another coefficient on the multi-profile case axis, the former also tends to be located above the latter on the single-profile case axis meaning that the ranking of relative magnitudes is preserved. The two exceptions are "public hospital" and to a much lesser extent "well staff", which are relatively more valued in the multi-profile case than in the single-profile case; however, "public hospital" is one of the less important job aspects. Overall, the more important coefficients are consistently ranked across methods meaning that the major policy recommendations would be similar across methods. Salary is again the exception: it is the second most important aspect in the multi-profile case, and yet its ranking drops to the fifth place in the single-profile case.

[Figure 4 about here]

Just as this is the first study of temporal stability in preferences elicited using the single-profile case, it provides the first evidence on the relative temporal stability

[21] Recall that our joint specification brings together two distinct behavioral models, the max-diff model for the single-profile case and the heteroskedastic rank-ordered logit model for the multi-profile case.

across the two experiment types. The bottom panel of Figure 4 plots between-wave shifts in the single-profile case coefficients (column 2 in Table 2) against between-wave shifts in multi-profile case coefficients (columns 4 in Table 2). The results suggest that intercase discrepancies are mostly preserved across waves. When it comes to non-salary coefficients, the between-wave shifts are fairly small and also there is no systematic pattern which suggests that identified intercase discrepancies in wave 1 are reinforced or mitigated over time; in other words, there is no discernible relationship between the upper and lower panels of Figure 4. Salary, again, stands out as an outlier. The bottom panel of Figure 4 shows that the downward shift in this coefficient on the single-profile case axis is much greater than the leftward shift on the multi-profile case axis. This means that the relative undervaluation of salary in the single-profile case in wave 1, as shown in the upper panel of Figure 4, has become even more apparent in wave 2. As mentioned above and shown in the right-most column of Table 2, the temporal shift in salary is the only one to differ significantly across the two methods at the 1% level.

Before moving to the heterogeneous preference model, we present some sensitivity checks to the choice of specification for salary. As discussed above, salary is an outlyer when evaluating the temporal and intercase stability of preference weights; given this, it is important to ensure that the use of a log-linear specification for salary has not affected our main conclusions. A likelihood ratio test leads to a rejection of the model with log-linear salary in favor of a dummy variable specification (the χ^2 value is 49.5 with 8 degrees of freedom). However, none of our conclusions are affected by the treatment of salary as a continuous variable. Appendix Figure A1 shows why. In that figure, we plot the preference weights over the relevant salary range for both the dummy variable and the log-linear salary specifications for both types of experiments. The figure shows that there is not quite enough curvature in the log function to capture the non-parametric estimates of the salary weights.[22] Nevertheless, the end points are very well captured by the log function and hence so are the shifts in preference weights over the range of salary values. For example, for wave 1, the increase in salary over the full range of values is 56% (1250/800). Based on the dummy variable specification, preference weights increase over this salary range by 4.260 in the single-profile case and 1.097 in the multi-profile case. Based on the log-linear salary specification, the corresponding figures are 4.011 and 1.096. As Appendix Figure A2 shows, the other preference weights

[22]The figure also shows why the log-linear specification performs better than the linear salary specification.

are also virtually unaffected by the choice between the log-linear salary specification and the salary dummies specification.[23]

4.2 Heterogeneous preference model

We now turn to our preferred model with unobserved preference heterogeneity. This framework approximates preference heterogeneity by allowing for four distinct preference segments or classes. The model is estimated by setting $C = 4$ in equation (6). As usual in the application of this type of "latent class" model, we have used the Bayesian Information Criterion (BIC) to choose the number of classes.[24] Tables 3 and 4 present results for this heterogeneous preference model, in a similar format to the homogeneous preference model reported in Table 2. The class shares reported at the bottom of each table range from 21 to 29%, and suggest that each class makes up roughly the same proportion of the population; in fact, we cannot reject the hypothesis that each share is 25%, both individually and jointly. We note that while the results have been split into two tables (Table 3 for the single-profile case and Table 4 for the multi-profile case) for the ease of presentation, all the results come from estimating one joint model as described in Section 3.

Beginning with the single-profile case results for wave 1 in Table 3, it is evident that the relative valuation of different job aspects vary from class to class, though supportive management and excellent quality of care always make the list of the three most important non-pecuniary attributes. Class 1 and Class 4 are more sensitive to salary than most of the other job attributes, whereas the reverse is true for Class 2 and Class 3. Class 2 and Class 3 have more comparable rankings of attributes, but Class 2 cares relatively more about supportive management whereas Class 3 cares more about the hospital's reputation for excellent quality of care. When comparing Class 1 to Class 4, we observe that Class 4 cares about three clinical rotations almost as much as supportive management, though for Class 1 the number of clinical rotations is the second least important attribute. The attribute "clinical rotations" is one of the few without a clear "better" level. In general, we would expect some students to prefer greater specialization while others would choose a broader training. The heterogeneity in preference weights for this attribute confirms this. The policy implication is that

[23] Detailed estimation results for the model with salary dummies are available upon request.

[24] Specifically, we have repeatedly estimated equation (6) by varying the number of preference segments from $C = 2$ to $C = 7$. The resulting BIC profile was U-shaped with BIC values of 33628 at $C = 2$ and 33819 at $C = 7$, reaching the minimum of 33381 at $C = 4$.

contracts should recognize this preference heterogeneity and offer choices over the degree of specialization to junior nurses.

[Table 3 about here]

On the issue of temporal stability, we find that all the main features of preference heterogeneity across classes remain robust over time. Our earlier findings on temporal stability from the homogeneous preference model is qualitatively replicated here for each class. There are statistically significant shifts in utility weights on some job characteristics over time (especially the middle ranked attributes), but the rankings of the attributes within each class are remarkably stable over time. To illustrate this point, Figure 5 provides the wave-by-wave comparison of utility weights across the four classes. The similarity of the upper panel for wave 1 and the lower panel for wave 2 is striking, especially when one focuses on the more important attributes located towards top on the vertical axis. For all classes, the utility weight on salary exhibits a large drop relative to changes in the utility weights on non-pecuniary attributes. The size of the drop is proportionally similar across all classes, ranging from 27% in class 2 to 32% in class 3 when compared to the level in wave 1, though it is only significant at the 1% level for Classes 1 and 4 the types that are more sensitive to salary.

[Figure 5 about here]

Table 4 reports estimation results for the multi-profile case, which closely resemble the single-profile case results in Table 3. For example, Classes 1 and 4 care relatively more about salary, Class 3 cares more about quality of patient care and Class 2 cares about supportive management. Although we see some attenuation of the strength of preferences over time, few shifts across the waves are significant. This can be also seen from Appendix Figure A3 that displays the multi-profile case results by wave; the wave 1 estimates and the wave 2 estimates show similar patterns across attribute levels and classes in that figure. Three of the 4 classes show a drop in the weight on salary although only one of them (Class 1) is significant.

[Table 4 about here]

The similarity of the preceding two tables and figures suggests that there is a good deal of intercase stability of preferences across all classes. To illustrate this point more explicitly, Figure 6 plots "demeaned" coefficients across different classes for each case of BWS in wave 1. These demeaned coefficients are derived by subtracting the population mean coefficients, reported in the last columns of Tables 3 and 4, from the corresponding class-specific coefficients; thus, a positive value (negative) indicates an above-average (below-average) utility weight on the relevant attribute.[25] The patterns across classes are similar regardless of whether one looks at the results for the single-profile case in the upper panel or the multi-profile case in the lower panel, and suggests that if a class has an above-average (below-average) utility weight on a certain attribute in one task, it also tends to have an above-average (below-average) utility weight on the same attribute in the other task. The exception is Class 1, which places relatively more weight on salary in the multi-profile case task. Finally, although we do not show a separate graph for this, it is evident from Tables 3 and 4 that the latent class results suggest a larger drop in the salary weight over time for the single-profile case, similarly to what was observed for the homogeneous preference model.

[Figure 6 about here]

4.3 Stability of stochastic error components

The preceding analysis has focused on the stability of preferences for attribute-levels, based on the comparison of utility weights across decision tasks and survey waves. In relation to the classic decomposition of a random utility into the systematic component that depends on attribute-levels and the unsystematic component that is modeled as a random disturbance, our focus has been on the stability of the systematic component. But considering that identified coefficients in any non-linear discrete choice model are inversely proportional to the latent error variance, it is possible that a change in the error variance, i.e. instability of the unsystematic component, is confounded with a genuine shift in the coefficient, i.e. instability of the systematic component. In fact, a study by Hensher *et al.* (1999) finds that sometimes, shifts in several coefficients

[25] Each population mean coefficient is derived by taking the weighted average of the four class-specific coefficients, with class shares used as weights.

across two sources of data can be modeled much more parsimoniously by introducing one parameter that captures the variance ratio between the two sources of data.[26]

In this section, we explore this alternative route of modeling temporal and intercase variations in coefficients as the implications of variations in the latent error variances. The goal is to understand how well this parsimonious approach would have explained the observed choices, relative to our general approach of allowing for shifts in individual coefficients separately. Note that it is not possible to identify a model that accounts for both shifts in the latent error variances and unrestricted shifts in all coefficients across the same dimensions; a change in the error variance implies changes in all coefficients by the same proportion, and these testable implications will be completely absorbed as part of the unrestricted coefficient shifts. To have any hope of detecting shifts in the error variances, it is essential to assume that at least a subset of coefficients remain stable across the dimensions of interest once the model controls for the shifts in the variances.

We consider two models that impose some form of coefficient stability *a priori* to identify different error variances in the four sources of data that we jointly analyze. We call the first model the "stable preference" model; it constrains all coefficients to be stable across the two cases of BWS and the survey waves, apart from potential re-scaling due to the variance shifts. We call the second model the "hybrid" model; it constrains coefficients on non-salary attributes to be stable as in the stable preference model, but allows coefficients on salary to vary in an unrestricted fashion as in our main analysis.

Appendix Table A1 compares selected results across the unrestricted model from Section 4.1 and its "stable preference" and "hybrid" special cases. Moving from the stable preference model to the hybrid model entails 4 extra parameters to relax the stability constraints on the salary coefficients, and improves the log-likelihood by 191.78. Moving from the hybrid model to the unrestricted model entails 40 extra parameters to relax the stability constraints on the non-salary coefficients, but improves the log-likelihood by a smaller amount of 109.68. This agrees with our conclusions above that

[26]In the context of the homogeneous preference model above, this means that the researcher may consider a more parsimonious model that takes the single-profile case coefficients in wave 1 as primitives and includes *three* scale parameters to account for proportionate shifts in the primitive coefficients across decision tasks and survey waves. The first scale parameter is for the variance ratio between the single-profile case in wave 1 and the multi-profile case in wave 1; the second one is for the variance ratio between the single-profile case in wave 1 and the single-profile case in wave 2; and the final one is for the variance ratio between the multi-profile case in wave 1 and the multi-profile case in wave 2.

preferences for salary vary more across methods and waves than preferences for other attributes. However, likelihood ratio tests reject both special cases in favor of the unrestricted model.

Appendix Table A1 also reports ratios of estimated error variances and tests results of the null of equal variances. There is limited evidence that temporal variations in the utility weights can be characterized as the upshot of shifting error variances. For the single-profile case, the ratio of variances in the two waves is very close to 1 and the null cannot be rejected in either of the two restricted models. This accords with the top panel of Figure 3 which shows that the slope of the trendline is very close to 1. For the multi-profile case, there is more evidence of an increase in the variance in the error over time; however, this shift over time is only significant in the hybrid model and even then it is marginally significant at the 10% level. There is however evidence that the shifts in the error variances may be a useful way to characterize the huge difference between the magnitudes of the single-profile case coefficients and the multi-profile case coefficients. The error variances are very different across the two tasks, with the variance of the error in the multi-profile case being around 50 times that of the single-profile case yielding a scaling factor of approximately 0.14 for multi-profile coefficients relative to single-profile coefficients. One interpretation of these results is that respondents indeed find the single-profile case task easier to complete, such that their choices are less influenced by random behavioral errors.

5 Conclusion

In this paper we compare stated preferences over job characteristics in entry-level nursing positions across time and across two types of Discrete Choice Experiments (DCEs), a traditional multi-profile case task and a recently proposed single-profile case task. We also develop and implement an econometric framework that combines the choices over time and across the two elicitation methods. Overall we find a surprising level of stability across time in the stated job preferences for the two types of DCEs. The similarity of preferences across methods is also comforting and lends support for the reliability of the stated preference methods.

There is an exception to this stability of preference weights and it regards the valuation of salary; namely, we find a marked drop in the weight on salary over time relative to other job attributes. Since this result is found for both types of tasks, it is

difficult to attribute it to a framing effect. The same group of individuals weighs salary less relative to other job characteristics 15 months later in similar experiments. As far as we are aware, the finding of a drop in the preference weights for salary over time relative to other job characteristics is new. Given the importance of tenure-wage profiles in the study of mobility and the provision of incentives, such a finding deserves further research. It points to the greater importance that should be given to non-salary job attributes and their profiles over time when designing labor contracts. The analysis of later waves of the underlying survey will shed light on whether the drop in the preference weight on salary continues or is only an initial feature of intertemporal preferences. More generally, it would be interesting to see if this holds for other occupations as well.

In previous work, we had found an undervaluation of salary in the new single-profile method relative to other preference weights. This result holds in this paper for each wave of the survey. We surmised in our previous paper that this discrepancy could be due a more direct comparison of salary with other attributes some of these with social connotations (e.g. quality of patient care) in the single-profile case. This speaks to a potential important framing effect in cases where respondents are asked to compare directly "socially valued attributes" with their own private benefit (salary). We are currently designing additional experiments to test this hypothesis.

We argued earlier that stated preferences can generate a more complete picture of job preferences by allowing the estimation of valuations over attributes that vary little in the market. An additional advantage illustrated in this paper is that stated preference methods can provide early indications of mismatch between relative preference weights over job characteristics and the design of actual jobs. In the case of a regulated job market such as nursing, often the only variation observed in labor market participants' behavior comes in the form of quitting and moving to different occupations (Frijters *et al.*, 2007). The results here suggest that policies and regulations regarding nursing jobs need to improve non-salary job conditions even for very junior nurses. For most junior nurses, management style is more important than quality of care, and given the heterogeneity in preferences over the degree of specialization, flexibility in the number of rotations should be allowed.

The ultimate proof of the validity and usefulness of these stated preference methods rests with their predictive power and hence with a comparison of stated and revealed preferences. Although we aim to use the survey on nursing students and new graduates to address this question, more waves of the survey must be utilized as we need enough variation in revealed preferences and this is left for further research. More generally, we

hope that a better understanding of the methodology underlying the data collection and analysis of stated preferences will encourage labor market researchers to consider the use of stated preferences over jobs either as stand alone information or in combination with usual survey questions.

References

Adamowicz V. 2013. Choice modelling and environmental valuation: Where have we been, what's up ahead? *Keynote address at: 2013 International Choice Modelling Conference, Sydney, Australia.* `http://www.icmconference.org.uk/index.php/icmc/ICMC2013/paper/viewFile/785/350` (last accessed: 07 September 2017).

Akaichi F, Nayga RM, Gil JM. 2013. Are results from non-hypothetical choice-based conjoint analyses and non-hypothetical recoded-ranking conjoint analyses similar? *American Journal of Agricultural Economics* **95**: 946–963.

Araújo EC, Maeda A. 2013. How to recruit and retain health works in rural and remote areas in developing countries: a guidance note. HNP Discussion Paper 78506. The World Bank.

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

Bhat C. 1997. An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* **31**: 34-48.

Blaauw D, Erasmus E, Pagaiya N, Tangcharoensathein V, Mullei K, Mudhune S, Goodman C, English M, Lagarde M. 2010. Policy interventions that attract nurses to rural areas: a multicountry discrete choice experiment. *Bulletin of World Health Organization* **88**: 350-356.

Brownstone D, Bunch D, Train K. 2000. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research B* **34**: 315–338.

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

Caparros A, Oviedo JL, Campos P. 2008. Would you choose your preferred option? Comparing choice and recoded rankingexperiments. *American Journal of Agricultural Economics* **90**: 843–855.

Czajkowski M, Bartczak A, Budzinski W, Giergiczny M, Hanley N. 2014. Within- and between- sample tests of preference stability and willingness to pay for forest management. Working Papers No. 24/2014 (141), University of Warsaw.

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

Doiron D, Hall J, Kenny P, Street D. 2014. Job preferences of students and new graduates in nursing. *Applied Economics* **46**: 924-939.

Doiron D, Yoo HI. 2017. Temporal stability of stated preferences: the case of junior nursing jobs. *Health Economics* **26**: 802-809.

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

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

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

Frijters P, Shields MA, Price SW 2007. Investigating the quitting decision of nurses: Panel data evidence from the British National Health Services. *Health Economics* **16**: 57-73.

Hole AR. 2006. CLOGITHET: Stata module to estimate heteroscedastic conditional logit model. Statistical Software Components S456737, Boston College Department of Economics, revised 09 Feb 2009.

Global Health Workforce Alliance (GHWA), World Health Organization (WHO). 2013. A Universal Truth: No Health Without a Workforce. *Third Global Forum on Human Resources for Health Report.*

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

Heckman J, Singer B. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**: 271-320.

Hensher DA, Louviere J, Swait J. 1999. Combining sources of preference data. *Journal of Econometrics* **89**: 197-221.

Holte JH, Kjaer T, Abelsen B, Olsen JA. 2015. The impact of pecuniary and non-pecuniary incentives for attracting young doctors to rural general practice. *Social Science and Medicine* **128**: 1-9.

Huicho L, Miranda JJ, Diez-Canseco F, Lema C, Lescano AG, Lagarde M, Blaauw D. 2012. Job Preferences of Nurses and Midwives for Taking Up a Rural Job in Peru: A Discrete Choice Experiment. *PLOS ONE* **7**: 1-9.

Islam T, Louviere J. 2015. The stability of aggregate-level preferences in longitudinal discrete choice experiments. In: Louviere J, Flynn T, and Marley A (eds.) *The Best-Worst Scaling: Theory, Methods and Applications*, Cambridge University Press, 265-277.

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

Krucien N, Watson V, Ryan M. 2016. Is best-worst scaling suitable for health state valuation? A comparison with discrete choice experiments. *Health Economics*: DOI: 10.1002/hec.3459.

Kolstad JR. 2011. How to make rural jobs more attractive to health workers: findings from a discrete choice experiment in Tanzania. *Health Economics* **20**: 196-211.

Layton D, Brown G. 2000. Heterogeneous preferences regarding global climate change. *Review of Economics and Statistics* **82**: 616-624.

Liebe U, Meyerhoff J, Hartje V. 2012. Test-retest reliability of choice experiments in environmental valuation. *Environmental and Resource Economics* **53**: 389-407.

Louviere J, Flynn TN, Marley AAJ. 2015. Best-worst Scaling: Theory, Methods and Applications. Cambridge University Press: Cambridge, UK.

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

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

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

McFadden D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In: Paul Zarembka (ed.), *Frontiers of Econometrics*. New York: Academic Press.

Pedersen LB, Gyrd-Hansen D. 2014. Preference for practice: a Danish study on young doctors' choice of general practice using a discrete choice experiment. *European Journal of Health Economics* **15**: 611-621.

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

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

Rigby D, Burton M, Pluske J. 2015. Preference stability and choice consistency in discrete choice experiments. *Environmental and Resource Economics*: forthcoming. DOI 10.1007/s10640-015-9913-1.

Ryan M, Netten A, Skatun D, Smith P. 2006. Using discrete choice experiments to estimate a a preference-based measure of outcome - An application to social care for older people. *Journal of Health Economics* **25**: 927-944.

Ryan M, Kolstad JR, Rockers PC, Dolea C. 2012. How to conduct a discrete choice experiment for health workforce recruitment and retention in remote and rural areas: a user guide with case studies. The World Health Organization, ISBN: 9789241504805.

Salkeld G, Solomon M, Butow P, Short L. 2005. Discrete choice experiment to measure patient preferences for the surgical management of colorectal cancer. *British Journal of Surgey* **92**: 742-747.

San Miguel F, Ryan M, Scott A. 2002. Are preferences stable? The case of health care. *Journal of Economic Behavior & Organization* **48** 1-14.

Schaafsma M, Brouwer R, Liekens I, De Nocker L. 2014. Temporal stability of preferences and willingness to pay for natural areas in choice experiments: A test-retest. *Resource and Energy Economics* **38**: 243-260.

Sivey P, Scott A, Witt J, Joyce C, Humphreys J. 2012. Junior doctors' preferences for specialty choice. *Journal of Health Economics* **31**: 813-823.

Skjoldborg S, Lauridsen J, Junker P. 2009. Reliability of the discrete choice experiment at the input and output level in patients with rheumatoid athritis. *Value in Health* **12**: 153-158.

Small KA, Winston C, Yan J. 2005. Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability. *Econometrica* **73**: 1367-1382.

Street DJ, Burgess L. 2007. The Construction of Optimal Stated Choice Experiments: Theory and Methods. Wiley, New York.

Street DJ, Burgess L, Louviere J. 2005. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing* **22**: 45970.

Street DJ, Knox S. 2012. Designing for attribute-level best-worst choice experiments. *Journal of Statistical Theory and Practice* **6**: 363–375.

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

Whitty JA, Ratcliffe J, Chen G, Scuffham. 2014. Australian Public Preferences for the Funding of New Health Technologies: A Comparison of Discrete Choice and Profile Case Best-Worst Scaling Methods. *Medical Decision Making* **34**: 638-654.

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

Yoo HI, Doiron D. 2013. The use of alternative preference elicitation methods in complex discrete choice experiments. *Journal of Health Economics* **32**: 1166-1179.

Figure 1: Sample single-profile case BWS scenario
Source: Yoo and Doiron (2013)

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 2: Sample multi-profile case BWS scenario
Source: Yoo and Doiron (2013)

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.

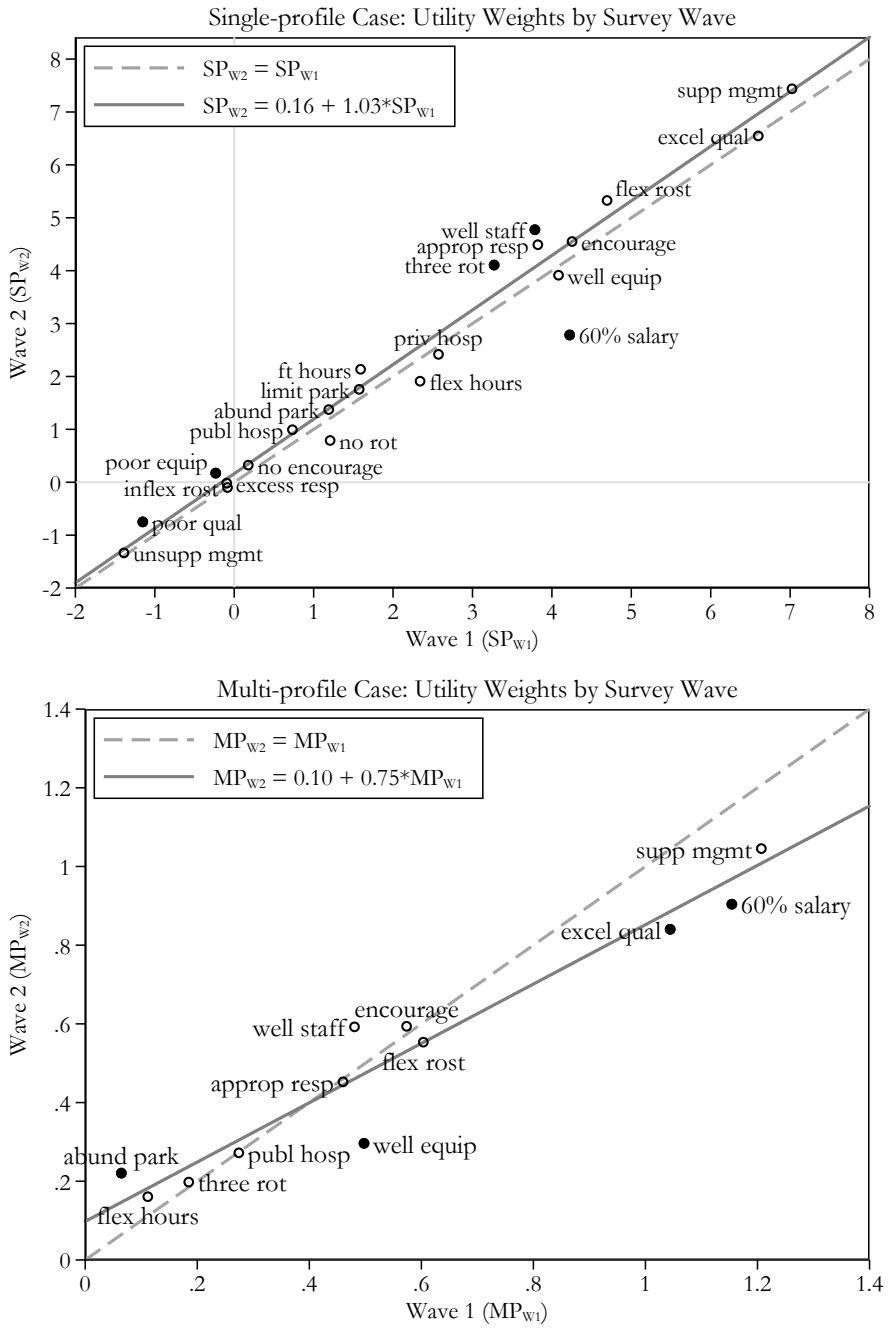
To review the features of jobs, please click here.

		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?

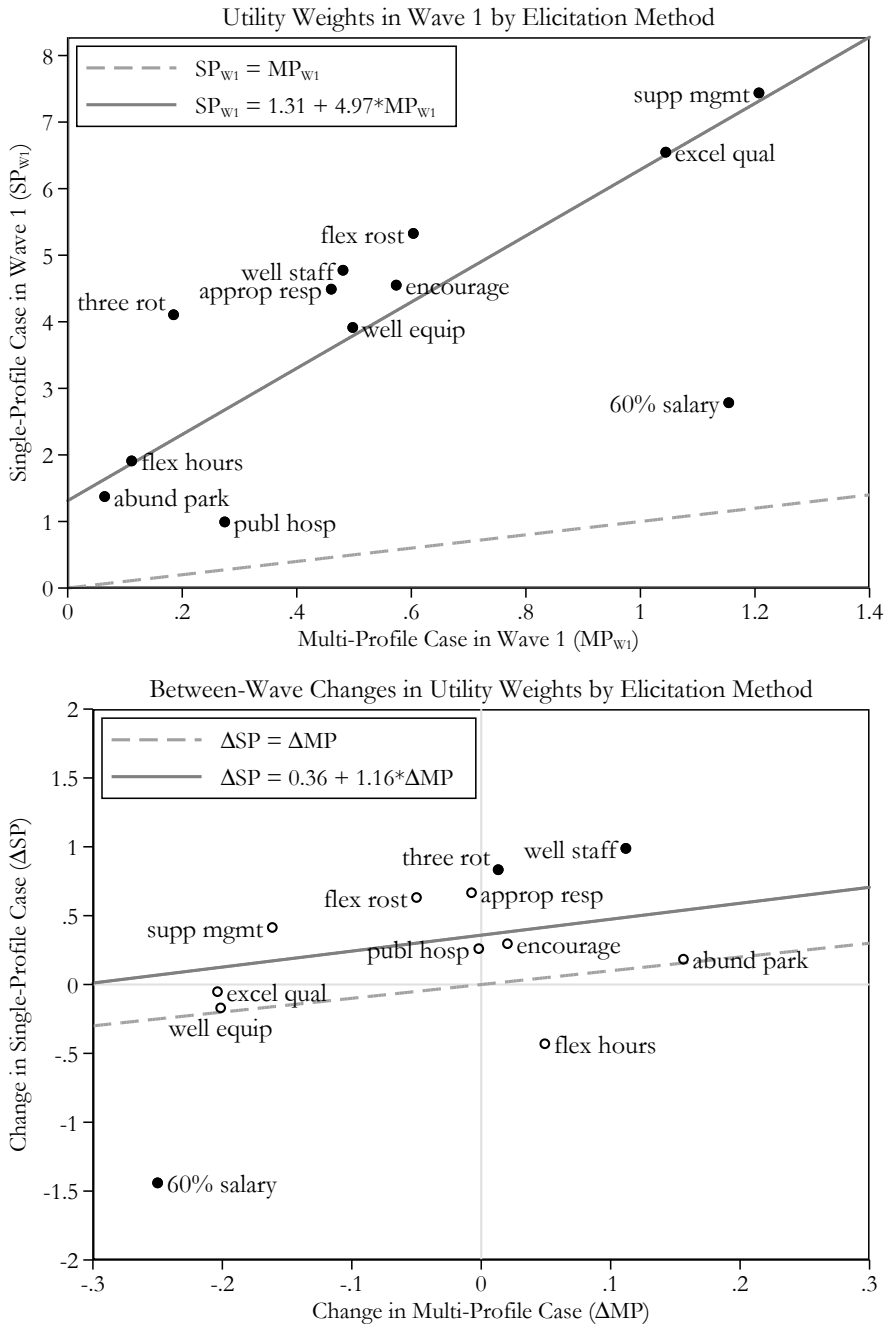
◯ Yes
◯ No

Figure 3: Homogeneous preference model results by survey waves



Notes: A filled (hollow) circle indicates that the pairwise difference is statistically significant (insignificant) at the 5% level.

Figure 4: Homogeneous preference model results by elicitation methods



Notes: A filled (hollow) circle indicates that the pairwise difference is statistically significant (insignificant) at the 5% level.

Figure 5: Heterogeneous preference model results for single-profile case

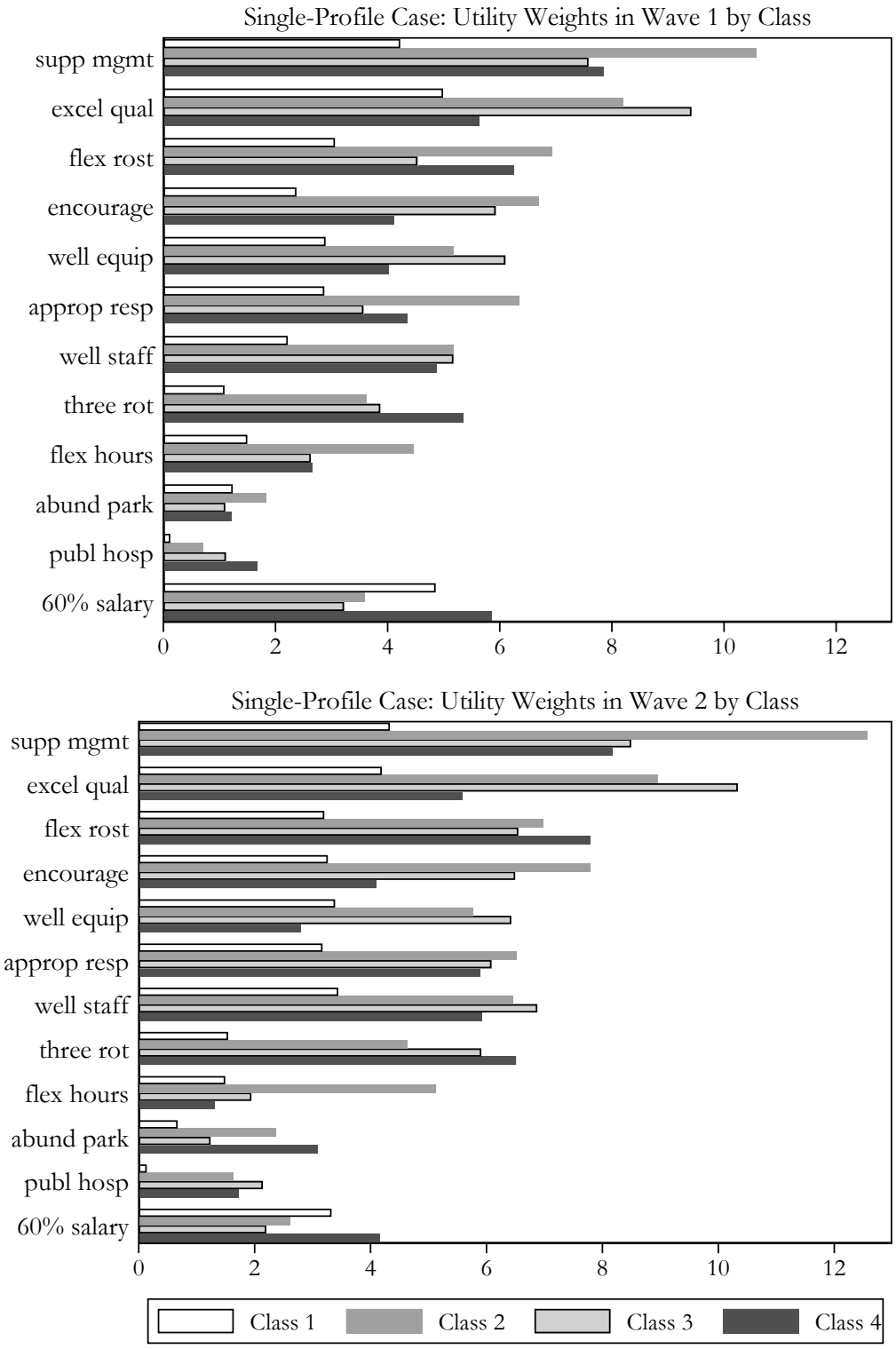


Figure 6: Heterogeneous preference model: demeaned results for wave 1

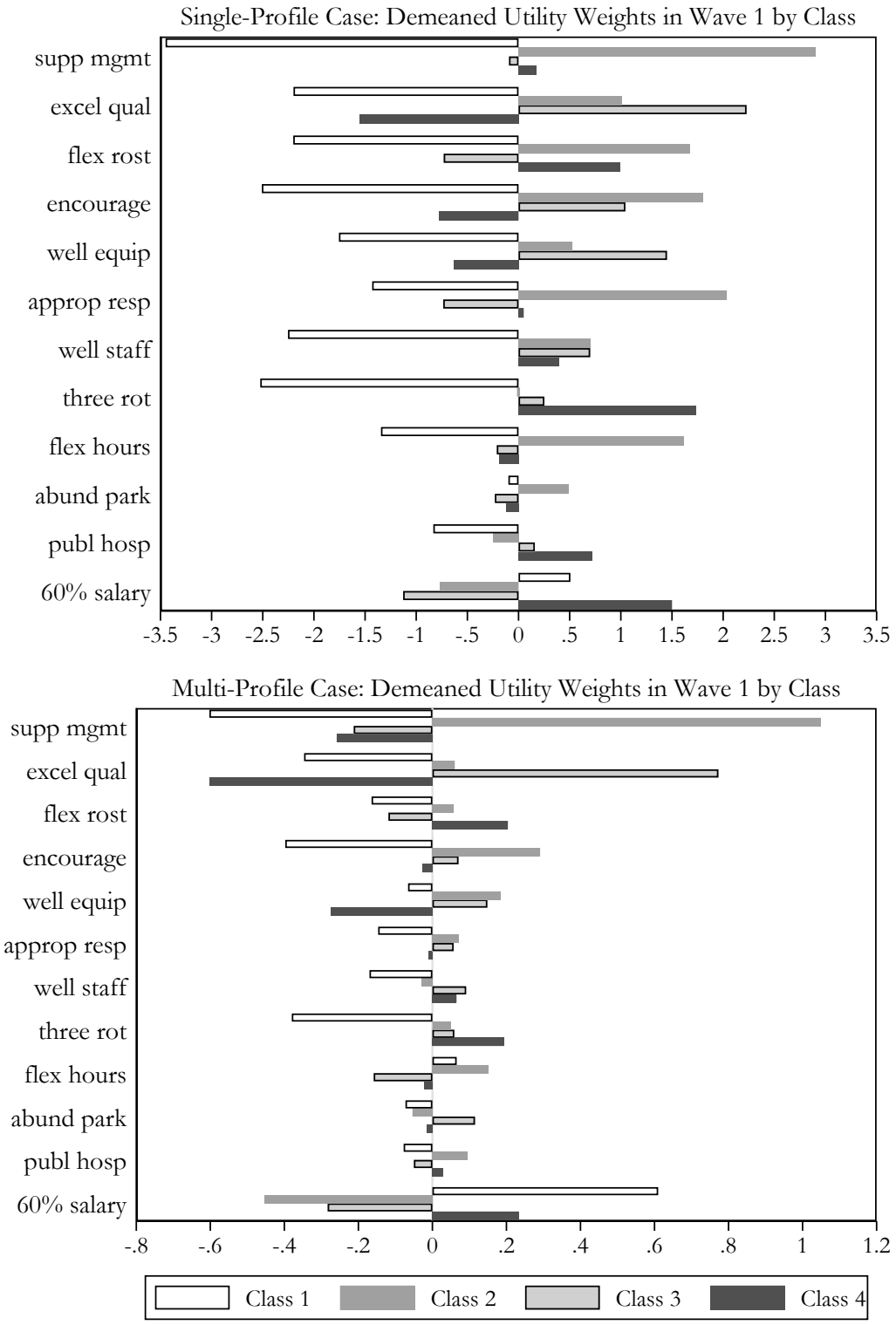


Table 1: Job attributes and associated levels

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	Wave 1: \$800, \$950, \$1100, \$1250 Wave 2: \$900, \$1100, \$1300, \$1500	ln(salary)

Table 2: Homogeneous preference model results

	A. Comparisons of identified coefficients across cases and waves					
	Single-profile (SP)		Multi-profile (MP)		SP − MP	
	W1	W2-W1	W1	W2-W1	W1	W2-W1
supportive mgt	7.024***	0.414	1.207***	-0.161*	5.810***	0.576
excellent care	6.599***	-0.052	1.044***	-0.204**	5.554***	0.152
flexible roster	4.695***	0.632*	0.603***	-0.050	4.092***	0.682*
encouraging env.	4.254***	0.296	0.574***	0.020	3.681***	0.276
well equipped	4.084***	-0.170	0.498***	-0.201***	3.586***	0.032
appropriate resp.	3.823***	0.667*	0.460***	-0.008	3.363***	0.674*
well staffed	3.786***	0.988**	0.481***	0.112*	3.305***	0.877**
three rotations	3.272***	0.834**	0.185***	0.013	3.087***	0.821**
flexible hours	2.340***	-0.430	0.112**	0.049	2.229***	-0.480
abundant parking	1.190***	0.184	0.064	0.156**	1.126***	0.028
public hospital	0.732***	0.261	0.274***	-0.002	0.458**	0.263
60% rise in salary	4.224***	-1.440***	1.154***	-0.250**	3.070***	-1.190***
	B. Coefficients identified in only one case					
private hospital	2.572***	-0.155				
full-time hours	1.591***	0.545*				
limited parking	1.573***	0.183				
no rotation	1.208***	-0.419*				
no encouragement	0.177	0.147				
inflexible roster	-0.085	-0.015				
excessive resp.	-0.095	0.077				
poorly equipped	-0.234*	0.406***				
poor care	-1.152***	0.403**				
unsupportive mgt	-1.389***	0.053				

*** indicates significant at 1%, ** at 5% and * at 10%, when the null hypothesis value is 1 for the σ parameter and 0 for all other parameters. All standard errors are clustered at the individual level. W1 reports estimates for wave 1, and W2-W1 reports wave 2-wave 1 differences in the estimates. All single-profile coefficients in panel A have been transformed to measure the extra utility that the shown level of each attribute offers over the omitted level of the same attribute, so that the results are comparable to the multi-profile case coefficients. The model involves 72 estimated parameters. The log-likelihood is -16875.88 for a sample of 234 individuals providing 3744 single-profile case responses (494,208 observations) and 3732 multi-profile case responses (18,660 observations).

Table 3: Heterogeneous preference model results for single-profile case

	A. Coefficients identified in all cases					
	Class 1		Class 2		Class 3	
	W1	W2-W1	W1	W2-W1	W1	W2-W1
supportive mgt	4.225***	0.108	10.580***	1.988**	7.583***	0.914
excellent care	4.990***	-0.795**	8.200***	0.758	9.418***	0.918
flexible roster	3.061***	0.141	6.936***	0.046	4.532***	2.017***
encouraging env.	2.375***	0.888**	6.689***	1.108	5.928***	0.567
well equipped	2.894***	0.494	5.176***	0.593	6.101***	0.327
appropriate resp.	2.872***	0.299	6.339***	0.179	3.567***	2.519***
well staffed	2.220***	1.223***	5.178***	1.273	5.174***	1.704***
three rotations	1.091***	0.451	3.615***	1.022	3.869***	2.042***
flexible hours	1.498***	-0.006	4.457***	0.674	2.630***	-0.686
abundant parking	1.238***	-0.566	1.826***	0.541	1.106**	0.134
public hospital	0.124	0.015	0.710	0.924	1.117**	1.025
60% rise in salary	4.858***	-1.531***	3.585***	-0.968	3.223***	-1.021*
Class shares	0.210		0.243		0.282	

	Class 4		Mean	
	W1	W2-W1	W1	W2-W1
supportive mgt	7.849***	0.325	7.676***	0.849***
excellent care	5.637***	-0.060	7.190***	0.259
flexible roster	6.255***	1.538***	5.263***	1.017***
encouraging env.	4.108***	-0.005	4.884***	0.614*
well equipped	4.016***	-1.221*	4.650***	0.017
appropriate resp.	4.354***	1.538***	4.303***	1.224***
well staffed	4.871***	1.045*	4.473***	1.324***
three rotations	5.352***	1.158**	3.616***	1.225***
flexible hours	2.655***	-1.343*	2.843***	-0.387
abundant parking	1.215**	1.869***	1.338***	0.545
public hospital	1.676***	0.048	0.957***	0.530
60% rise in salary	5.851***	-1.697***	4.351***	-1.294***
Class shares	0.265			

continued on next page

Table 3: *(continued)*

	B. Coefficients identified only in single-profile case					
	Class 1		Class 2		Class 3	
	W1	W2-W1	W1	W2-W1	W1	W2-W1
private hospital	2.235***	0.057	3.191***	-1.343*	2.968***	-0.911
full-time hours	1.452***	0.198	0.963***	0.298	2.692***	0.803
limited parking	0.647**	0.942**	1.969***	-0.563	2.335***	0.27
no rotation	1.112***	0.001	1.926***	-0.939	1.714***	-0.833*
no encouragement	0.47	-0.369	-0.43	-0.088	0.113	0.264
inflexible roster	-0.053	0.231	-0.509*	0.199	0.910***	-0.486
excessive resp.	-0.214	0.18	-0.714	0.504	1.210***	-1.066***
poorly equipped	-0.537**	0.401	-0.174	0.005	-0.339*	0.155
poor care	-1.166***	0.865***	-1.432	0.095	-1.741***	0.117
unsupportive mgt	-0.787***	0.156	-2.363	-0.933**	-0.922***	0.136

	Class 4		Mean	
	W1	W2-W1	W1	W2-W1
private hospital	2.798***	0.623	2.823***	-0.406
full-time hours	1.525***	1.539***	1.702***	0.748***
limited parking	2.386***	-0.668	1.905***	-0.040
no rotation	0.740***	-0.241	1.381***	-0.527**
no encouragement	0.578**	0.680*	0.179	0.156
inflexible roster	-0.490**	-0.090	-0.008	-0.064
excessive resp.	-0.383*	0.300	0.021	-0.061
poorly equipped	0.220	1.428***	-0.193*	0.508***
poor care	-0.236	0.748**	-1.146***	0.436***
unsupportive mgt	-1.501***	0.472*	-1.397***	-0.031

*** indicates significant at 1%, ** at 5% and * at 10%, when the null hypothesis value is 0.25 for class share parameters and 0 for all other parameters. These results have been jointly estimated with the multi-profile case results reported in Table 4. W1 reports estimates for wave 1, and W2-W1 reports wave 2-wave 1 differences in the estimates. Mean is a derived statistic and reports the weighted average of class-specific coefficients, using the class shares as weights. All single-profile coefficients in panel A have been transformed to measure the extra utility that the shown level of each attribute offers over the omitted level of the same attribute, so that the results are comparable to the multi-profile cae coefficients. The model involves 291 estimated parameters. The log-likelihood is -15897.058 for a sample of 234 individuals providing 3744 single-profile case responses and 3732 multi-profile case responses. All other information is as provided in Table 2.

Table 4: Heterogeneous preference model results for multi-profile case

	Class 1		Class 2		Class 3	
	W1	W2-W1	W1	W2-W1	W1	W2-W1
supportive mgt	0.734***	-0.086	2.386***	-0.490**	1.124***	-0.302*
excellent care	0.804***	-0.139	1.211***	-0.524***	1.923***	-0.348*
flexible roster	0.491***	-0.220	0.713***	-0.201	0.536***	0.233
encouraging env.	0.252**	0.043	0.941***	0.0436	0.720***	-0.072
well equipped	0.527***	-0.293*	0.777***	-0.304	0.742***	-0.224
appropriate resp.	0.336***	-0.044	0.553***	-0.0193	0.540***	0.144
well staffed	0.341***	0.161	0.482***	-0.001	0.603***	0.093
three rotations	-0.152	0.086	0.277**	-0.00575	0.287**	-0.133
flexible hours	0.218*	0.004	0.303***	0.133	-0.00649	0.109
abundant parking	0.0122	0.428***	0.0311	-0.0416	0.200*	0.045
public hospital	0.227*	-0.015	0.399***	-0.138	0.254**	0.119
60% rise in salary	1.780***	-0.425*	0.716***	-0.375	0.887***	-0.311
σ	0.484***	0.287*	0.645***	0.086	0.622***	-0.039
Class shares	0.210		0.243		0.282	

	Class 4		Mean	
	W1	W2-W1	W1	W2-W1
supportive mgt	1.078***	-0.034	1.336***	-0.231**
excellent care	0.548***	-0.004	1.150***	-0.255***
flexible roster	0.859***	-0.139	0.655***	-0.066
encouraging env.	0.623***	-0.051	0.650***	-0.015
well equipped	0.318***	-0.239	0.593***	-0.262***
appropriate resp.	0.472***	-0.039	0.482***	0.017
well staffed	0.577***	0.142	0.512***	0.097
three rotations	0.421***	0.101	0.228***	0.006
flexible hours	0.130	-0.171	0.152***	0.019
abundant parking	0.070	0.180	0.085	0.140*
public hospital	0.333***	-0.058	0.304***	-0.018
60% rise in salary	1.403***	0.019	1.170***	-0.263**
σ	0.705***	-0.014		
Class shares	0.265			

*** indicates significant at 1%, ** at 5% and * at 10%, when the null hypothesis value is 0.25 for class share parameters, 1 for the σ parameters and 0 for all other parameters. These results have been jointly estimated with the single-profile case results reported in Table 3. All other information is as provided in Table 3.

Online Appendix
Stated preferences over job characteristics: A panel study

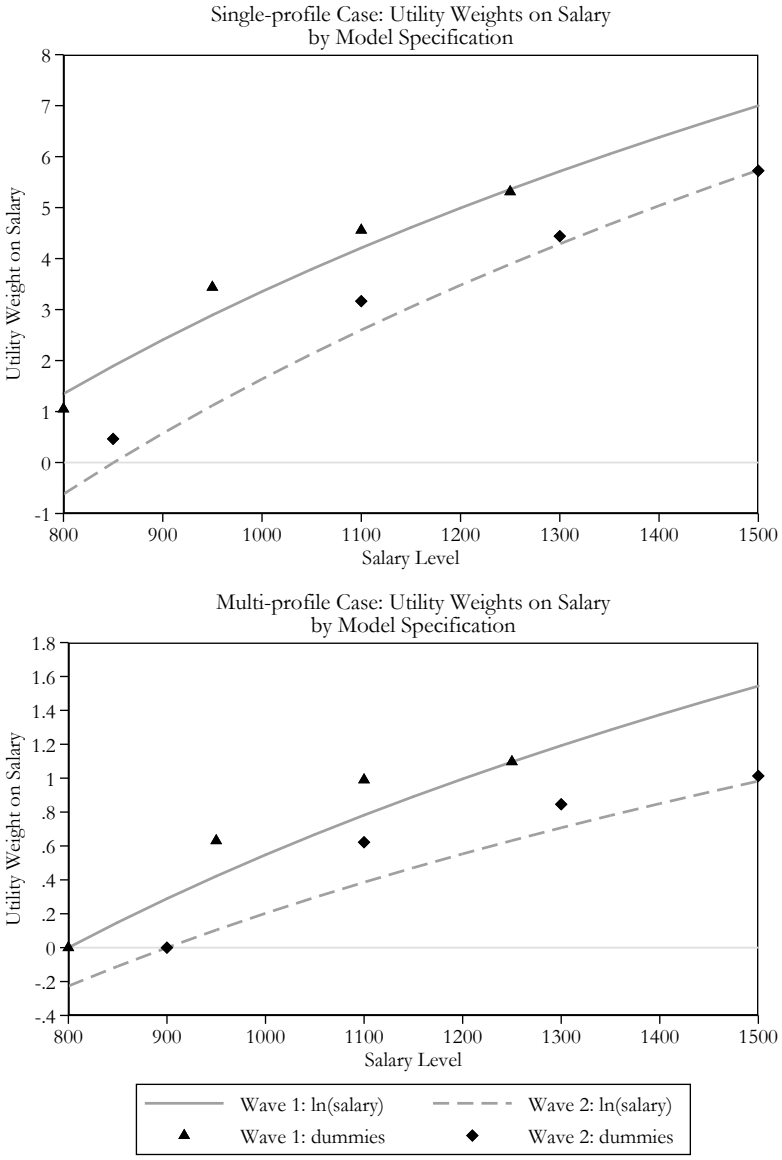
Denise Doiron[a]
Hong Il Yoo[b]

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

[b] Durham University Business School, Durham University, UK. h.i.yoo@durham.ac.uk

May 20, 2019

Figure A1: Sensitivity of utility weights on salary to model specifications



Notes: The $ln(salary)$ specification refers to our main specification that estimates the slope coefficients on $ln(salary)$; we plot the implied utility weights at relevant salary levels. The *dummies* specification refers to an alternative specification that replaces $ln(salary)$ with binary indicators of specific salary levels, thereby estimating utility weights at those levels directly.

Figure A2: Homogeneous preference model results by salary specifications

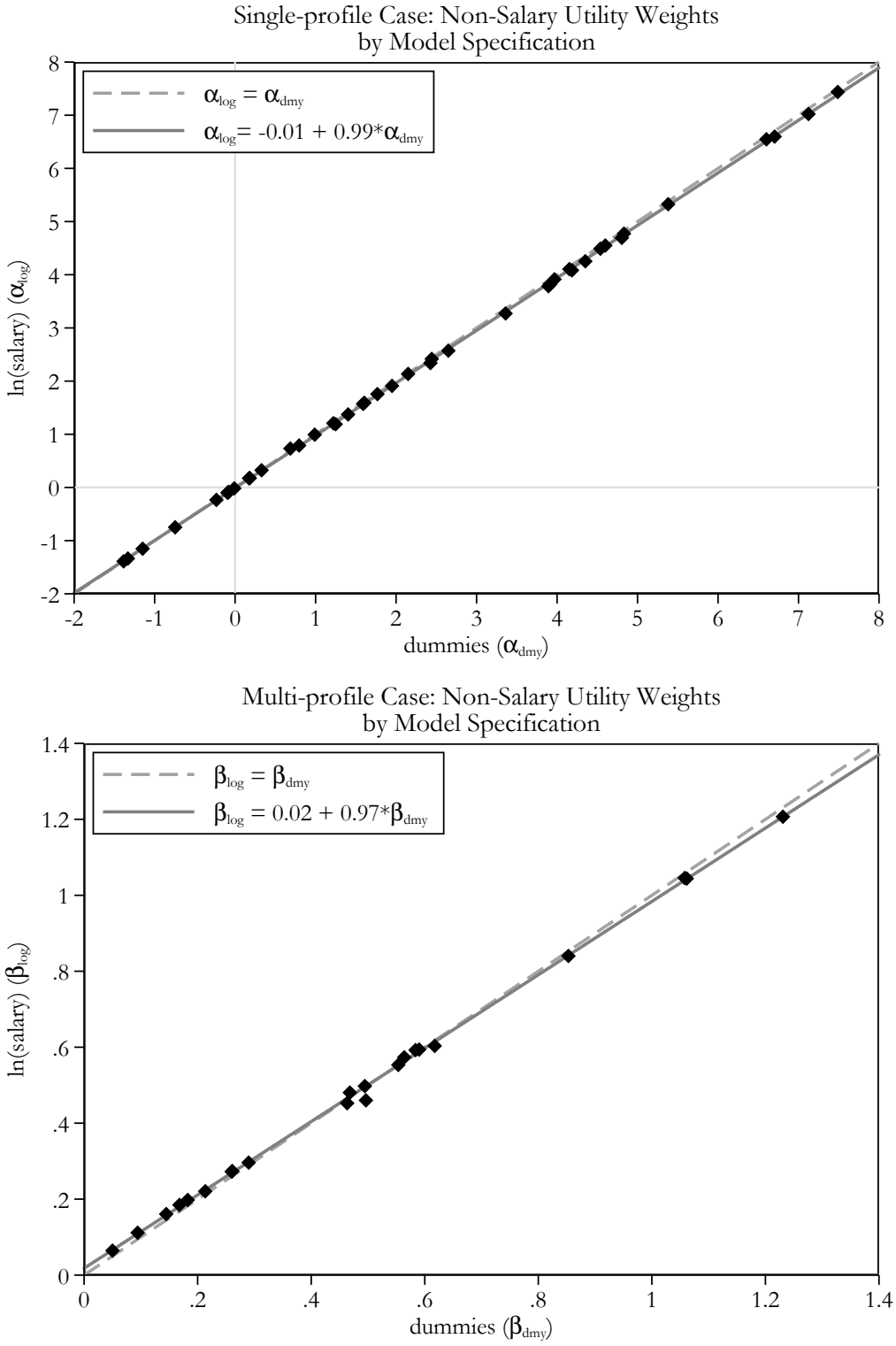


Figure A3: Heterogeneous preference model results for multi-profile case

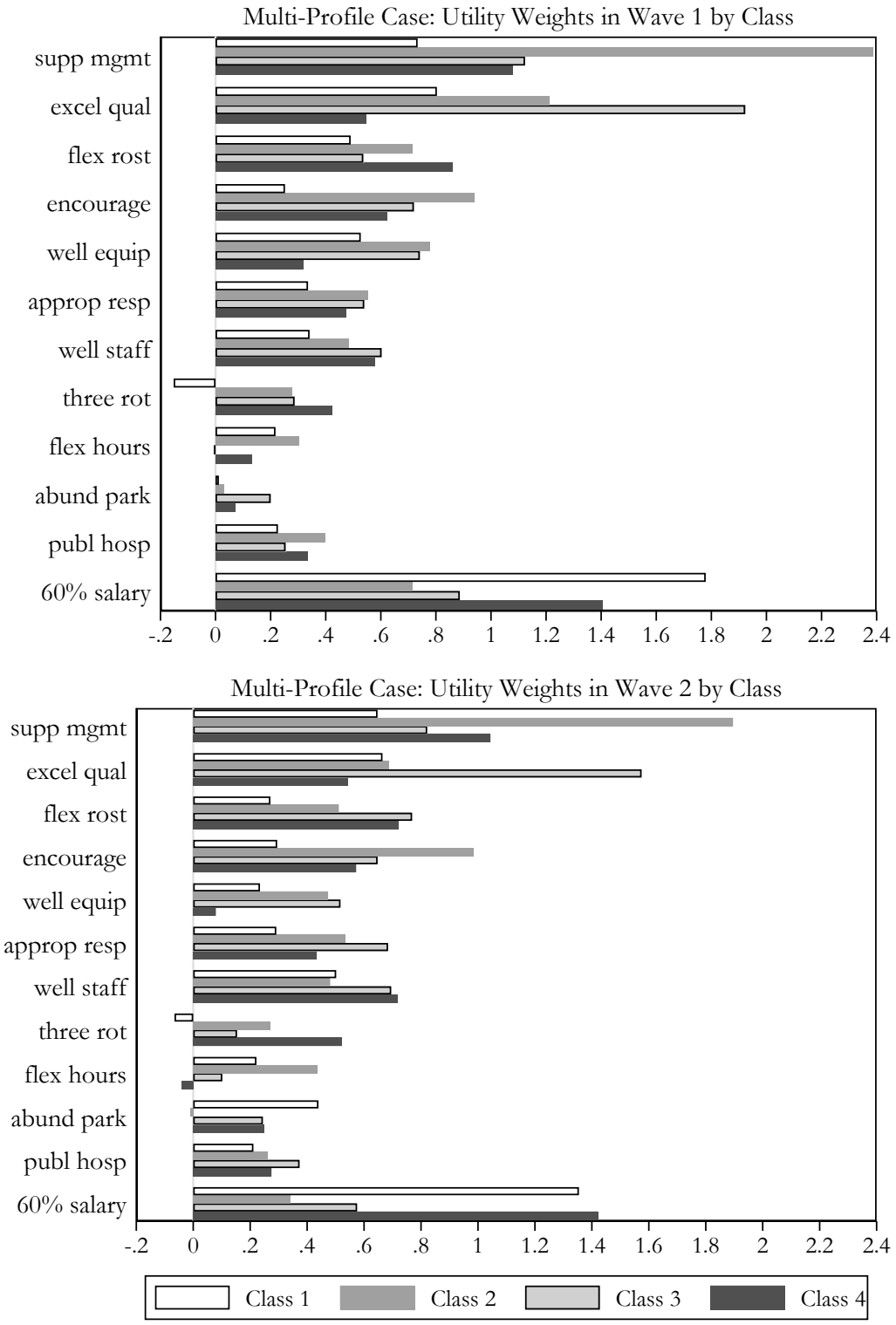


Table A1: Stability of latent error variances

A. Stable preference model			
Variance ratio	Value	$\chi^2(1)$	p-value
$Var(SP,2)/Var(SP,1)$	1.032	0.11	0.745
$Var(MP,2)/Var(MP,1)$	1.221	2.29	0.130
$Var(SP,1)/Var(MP,1)$	0.023	156971	0.000
$Var(SP,2)/Var(MP,2)$	0.020	192648	0.000
log-likelihood	-17177.34		
No. parameters	28		

B. Hybrid model			
Variance ratio	Value	$\chi^2(1)$	p-value
$Var(SP,2)/Var(SP,1)$	0.981	0.02	0.898
$Var(MP,2)/Var(MP,1)$	1.270	2.79	0.095
$Var(SP,1)/Var(MP,1)$	0.021	163998	0.000
$Var(SP,2)/Var(MP,2)$	0.016	242657	0.000
log-likelihood	-16985.56		
No. parameters	32		

C. Unrestricted model	
log likelihood	-16875.88
No. parameters	72

The unrestricted model is the homogeneous preference model in Table 2. The stable preference model is its special case that assumes stable coefficients across time and decision tasks but allows for unequal variances across time and decision tasks. The hybrid model allows preferences for salary to change freely across time and decision tasks, but is otherwise identical to the stable preference model. Var(SP,w) denotes the latent error variance of the max-diff model (for the single-profile case data) in wave w. Var(MP,w) denotes the latent error variance of the first-best component of the heteroskedastic rank-ordered logit model (for the multi-profile case data) in wave w. $\chi^2(1)$ reports the Wald test statistic for the null hypothesis of equal variances (i.e. the variance ratio equals one).