Temporal stability of stated preferences: the case of junior nursing jobs

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

a. School of Economics, University of New South Wales

Sydney, Australia. d.doiron@unsw.edu.au

b. Durham University Business School, Durham University

Durham, United Kingdom. h.i.yoo@durham.ac.uk

January 2016

JEL classification: C35, C81 , J44, I11

Key words: job preferences, discrete choice experiment, reliability, panel data, health workforce

Counts: 2000 words; 2 figures; 1 table.

Funding source: This work was supported by Discovery Project Grant DP0881205 from the Australian Research Council.

Conflict of interest: The authors do not have any conflict of interest.

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

Abstract

With the growing use of discrete choice experiments (DCEs) in health workforce research, the reliability of elicited job preferences is a growing concern. We provide the first empirical evidence on the temporal stability of such preferences, using a unique longitudinal survey of Australian nursing students and graduate nurses. The respondents completed DCEs on nursing positions in two survey waves. Each position is described by salary and eleven non-salary attributes, and the two waves are spaced 15 months apart on average. Between the waves, most final-year students finished their degrees and started out as graduate nurses. Thus, the survey covers a long timespan that includes an important period of career transition. The relative importance of different job attributes appears stable enough to support the use of DCEs to identify key areas of policy intervention. There is virtually no change in the groupings of influential job characteristics. Conclusions regarding the stability of willingness-to-pay, however, are different due to unstable preferences for salary. The instability of preferences for salary was also found previously in the context of comparing alternative elicitation methods. This prompts us to push for further work on the reliability of stated preferences over monetary attributes.

1 Introduction

Recent years have seen an increasing demand for empirical evidence to guide the recruitment and retention of health care professionals. Household surveys and registry data, however, do not often provide rich enough information to analyse trade-offs across a wide range of job attributes. A growing body of evidence comes from discrete choice experiments (DCEs), with the World Health Organization and the World Bank now promoting the DCE approach (Ryan *et al.*, 2012; Araújo and Maeda, 2013). This approach collects stated choices among job profiles to elicit preferences for job attributes. The Australian DCE study of Doiron *et al.* (2014), for example, is the first study from a developed country to provide evidence on trade-offs across several nursing job attributes, which have not been previously analysed together due to data limitations. Other recent DCE studies include Kolstad (2011) on Tanzanian clinical officers, Sivey *et al.* (2012) on Australian doctors, and Holte *et al.* (2015) on Norwegian doctors.

The usefulness of DCE studies depends on whether they capture fundamental aspects of preferences which can inform future decisions, making temporal stability an important reliability criterion. A handful of studies have tested for temporal stability explicitly. Earlier studies find stable preferences for medical services (Bryan *et al.*, 2000; San Miguel *et al.*, 2002; Salkeld *et al.*, 2005; Skjoldborg *et al.*, 2009) and social care outcomes (Ryan *et al.*, 2006) but, as Liebe *et al.* (2012) point out, their study designs may be conducive to carry-over effects since the repeat DCE occurs soon (2 weeks to 4 months) after the initial DCE. Recent studies in environmental valuation suggest that the timespan between DCEs indeed matters. Two six-month studies (Czajkowski *et al.*, 2014; Rigby *et al.*, 2015) support stability, while two one-year studies (Liebe *et al.*, 2012; Schaafsma *et al.*, 2014) present mixed evidence. The marketing study of Islam and Louviere (2015) on household consumables (e.g. toothpastes), however, finds that stability may hold over two years when DCE profiles describe familiar objects.

This paper is the first DCE study to investigate the temporal stability of stated job preferences. Exploring temporal stability over a long timespan is important in health workforce research, where DCEs often recruit prospective workforce cohorts 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. While prospective and current employees often have opportunities to experience various job attributes, job decisions are inherently more complex than the purchase of household consumables. Assuming temporal stability over a long period without relevant evidence may lead to erroneous policy decisions, especially when the period in question involves career transition. As Islam and Louviere (2015) review, behavioural and psychological research contends that preferences encompass constructed, as opposed to inherently stable, components that vary with external conditions. Qualitative research findings show that the health workers' assessment of job aspects varies within a few months into the first employment (Kelly and Ahern, 2009), suggesting that career transition may be one such condition. Whether job preferences remain stable, and if not, to what extent instability alters conclusions from an initial DCE, are questions that need be addressed empirically.

We analyse a longitudinal survey of nursing students and junior nurses who completed two waves of DCEs involving entry-level nursing positions in Australia.[1] The two waves were spaced at least a full year apart and 15 months on average, meaning that our study covers a longer timespan than all but one previous study (Islam and Louviere, 2015). Many of our respondents are making the transition from study to work, potentially learning a lot about performing nursing jobs.

2 Data and methods

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 in Australia.[2] This paper focuses on 241 respondents who participated in DCEs involving entry-level nursing jobs in two consecutive waves.[3] 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.

The first-wave DCE requires each respondent to complete 8 choice scenarios. Each scenario asks for the best and the worst out of 3 jobs simultaneously, thereby eliciting a

[1]Doiron *et al.* (2014) provide an in-depth analysis of the first wave of the survey.

[2]More information on this survey is available in Yoo and Doiron (2013) and Doiron *et al.* (2014).

[3]An earlier version of this paper provides a multinomial logit analysis of the current estimation sample, as well as an unbalanced sample that includes all 628 first-wave respondents. That version can be accessed at: `https://goo.gl/TFiSgJ`.

full preference ordering. The jobs are differentiated by weekly salary (four possible levels) and 11 non-salary attributes (two possible levels each): the online appendix provides details on the attribute-levels. The second-wave DCE is identical, barring two differences. First, the four underlying salary levels change from $\{\$800, \$950, \$1100, \$1250\}$ to $\{\$900, \$1100, \$1300, \$1500\}$, mirroring the updated pay scale for entry-level nursing jobs in 2011. Second, the respondents are not required to complete all 8 scenarios, although almost everyone (234 out of 241) voluntarily completed all.

Most DCEs prompt choices from available alternatives, instead of full preference orderings.[4] To obtain more directly comparable results, we focus on modelling which of 3 jobs is the best. In the random utility maximisation framework, the chosen and best alternatives are conceptually equivalent, since the decision maker is assumed to choose their best alternative. Capparos *et al.* (2008) and Akaichi *et al.* (2013) test this equivalence by administering the same DCE using both elicitation formats, and find supporting evidence.

Our analysis follows the usual random utility maximisation framework. The utility that person n derives from job j in choice senario t is

$$\begin{aligned} U_{njt} &= \boldsymbol{\beta}'_{nt}\boldsymbol{x}_{njt} + \varepsilon_{njt} \\ \text{where } \boldsymbol{\beta}_{nt} &= \boldsymbol{\alpha}_n + \boldsymbol{\delta}_n \times wave2_{nt.} \end{aligned} \tag{1}$$

$\boldsymbol{x}_{njt}$ is a vector of 12 attributes, $\boldsymbol{\beta}_{nt}$ is a conformable vector of utility coefficients, $wave2_{nt}$ is a dummy variable that equals 1 for choice scenarios from wave 2, and error term ε_{njt} is i.i.d. Type 1 Extreme Value. $\boldsymbol{\beta}_{nt}$ comprises $\boldsymbol{\alpha}_n$ that captures the baseline preferences in wave 1, and $\boldsymbol{\delta}_n$ that captures the deviations from the baseline in wave 2. We accommodate interpersonal taste heterogeneity nonparametrically by specifying a finite-mixture or "latent class" logit (LCL) model (Train, 2008): $\boldsymbol{\theta}_n = (\boldsymbol{\alpha}_n, \boldsymbol{\delta}_n)$ can be one of C classes or types of preferences, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_C$, with the population share of class c being $\Pr(\boldsymbol{\theta}_n = \boldsymbol{\theta}_c) = \pi_c$.[5] Our discussion focuses on the 3-class LCL model that estimates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \pi_1$, and π_2, normalising $\pi_3 = 1 - \pi_2 - \pi_1$.[6]

[4]To our best knowledge, Islam and Louviere (2015) is the only temporal stability study that models best-worst choices.

[5]The online appendix reports results based on other types of mixed logit models.

[6]The preferred number of classes C is often selected using the Bayesian Information Criterion (BIC). In our application, BIC for the 3-class model is 6260, compared to 6281 for the 2-class model and 6325 for the 4-class model.

All non-salary attributes are dummy-coded. Salary is log-transformed to allow for the diminishing marginal utility of money.[7]

3 Results

Figure 1 plots the second wave (W2) coefficients against the first wave (W1) coefficients. They are the population mean utility weights, $E(\boldsymbol{\alpha}_n + \boldsymbol{\delta}_n) = \sum_{c=1}^{3} \pi_c(\boldsymbol{\alpha}_c + \boldsymbol{\delta}_c)$ in W2 and $E(\boldsymbol{\alpha}_n) = \sum_{c=1}^{3} \pi_c \boldsymbol{\alpha}_c$ in W1. To facilitate interpretation, we multiply the coefficient on $\ln(salary)$ by $\ln(1.6)$, and plot the resulting utility weight on 60% extra salary: in W1 (W2), the largest salary level of \$1250 (\$1500) is 56% (67%) above the base level of \$800 (\$900).

Most coefficients are clustered around the 45° line, implying similar magnitudes over time. Some show marked deviations from it, however, especially the weights on better hospital equipment and 60% extra salary. Statistically, 5 out of 12 coefficients show significant changes, including these two. As environmental valuation studies with the longest timespan (Liebe *et al.*, 2012; Schaafsma *et al.*, 2014), we find evidence against the complete stability of utility weights.

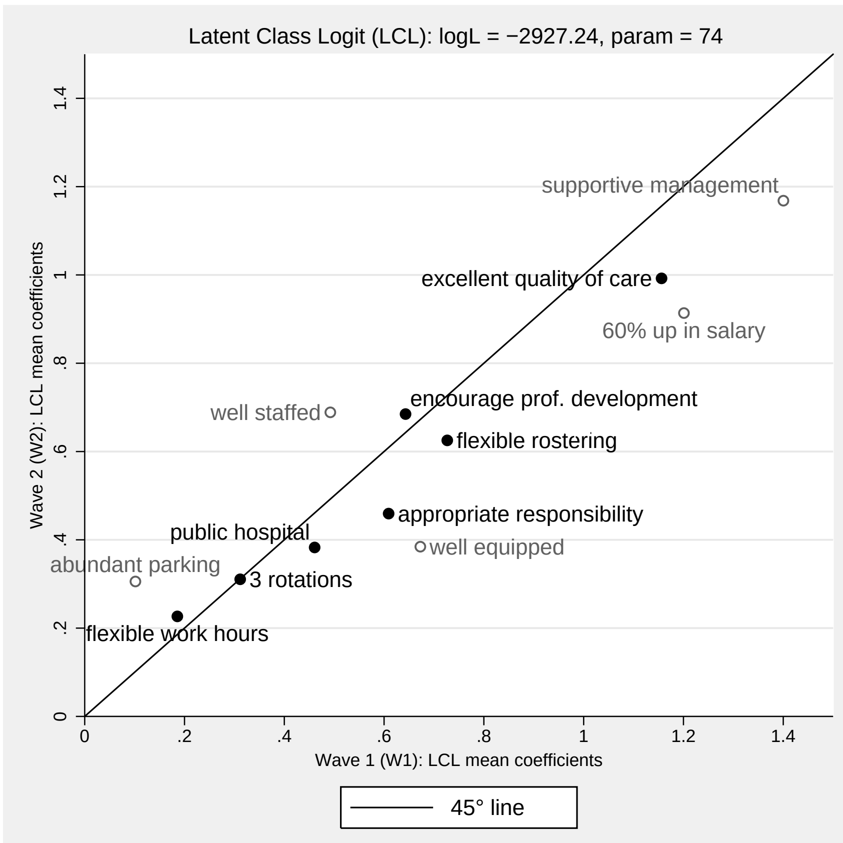
The relative valuation of job aspects, nevertheless, seems stable enough to allow its use in identifying the priority areas of intervention. Whether Figure 1 is read horizontally or vertically, attribute labels are encountered in a similar order, implying a similar relative valuation over time. The magnitude-based rankings of the coefficients in Table 1, indeed, show that 8 of 12 attributes display no or one-place change in rankings. The composition of four groups of three attributes based on W1 rankings remains unchanged in W2, except "well equipped" and "well staffed" that swap their membership in the 4th-6th place group and the 7th-9th place group. The results are particularly impressive considering that consistently trading off 12 attributes across 3 jobs is a non-trivial task.[8]

All money metric measures of better non-salary attributes can be expected to be influenced by the marked decline in the utility of extra salary in Figure 1. We consider

[7]The log transformation also adjusts for general price variations over person-specific completion dates. Any price deflator drops out from differences in the log of deflated salary within a choice set.

[8]Most other DCEs in health workforce research specify 7 attributes or less, albeit they tend to be more complex than ours in another dimension by specifying more than 2 levels per non-salary attribute: see for example Sivey *et al.* (2012) and Holte *et al.* (2015). Given the results on the effects of complexity found in DCEs involving health insurance (Barnes *et al.*, 2015) and environmental valuation (DeShazo and Fermo, 2002), we speculate that the stable relative valuation of job aspects can be generalised to health workforce DCEs involving fewer attributes than ours.

Figure 1: Wave-specific mean coefficients



The sample includes 3821 choice observations from 241 individuals. LogL (param) refers to the maximised log-likelihood (number of estimated parameters). The W1 coefficient on "abundant parking" is insignificant at the 10% level. All other coefficients are significant at the 1% level. A hollow circle indicates a significant between-wave difference at the 5% level. The base level for "public hospital" is "private hospital". Other non-salary attributes are vertical attributes and their base levels can be easily inferred. See the online appendix for further information on attribute-levels and detailed estimation results.

Table 1: Mean coefficients and magnitude rankings

	LCL Mean Coefficient			Ranking		
	W1	W2	W2−W1	W1	W2	W2−W1
supportive management	1.400***	1.168***	−0.232**	1	1	0
60% up in salary	1.201***	0.914***	−0.287**	2	3	+1
excellent quality of care	1.156***	0.992***	−0.164	3	2	−1
flexible rostering	0.727***	0.625***	−0.102	4	6	+2
well equipped	0.673***	0.384***	−0.289***	5	8	+3
encourage prof. development	0.643***	0.685***	+0.042	6	5	−1
appropriate responsibility	0.609***	0.459***	−0.150	7	7	0
well staffed	0.492***	0.689***	+0.197**	8	4	−4
public hospital	0.461***	0.383***	+0.078	9	9	0
3 rotations	0.312***	0.311***	−0.001	10	10	0
flexible work hours	0.186***	0.227***	+0.041	11	12	+1
abundant parking	0.102	0.306***	+0.204**	12	11	−1

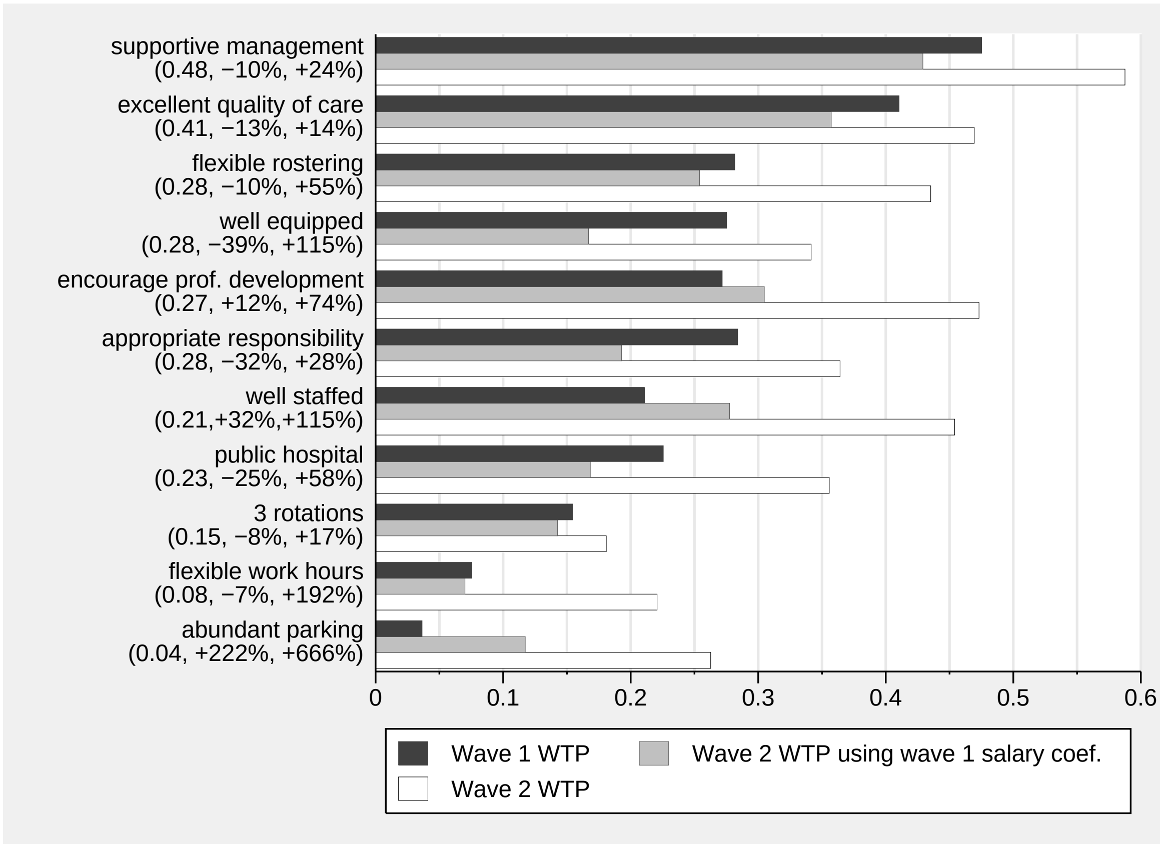
*, **, *** indicate statistical significance at the 10%, 5% and 1% levels. W1 and W2 stand for Wave 1 and Wave 2 respectively. Ranking refers to the rankings of the LCL mean coefficients in terms of magnitude.

the willingness-to-pay (WTP) or a fraction of salary that person n is willing to give up for an improvement in attribute k. The WTP equals $1 - \exp(-\theta_{nk}/\theta_{nS})$, given coefficients θ_{nk} on attribute k and θ_{nS} on $\ln(salary)$.[9] In the LCL model, the population mean WTP (MWTP) is obtained by averaging class-specific WTP using class shares as weights.

Figure 2 reports MWTP in each wave. We also compute MWTP for a counterfactual case that combines θ_{nk} from W2 and θ_{nS} from W1. As expected, MWTP in W2 tends to exceed MWTP in W1. But MWTP in the counterfactual case shows much less pronounced deviations, and suggest that the instability in MWTP is driven primarily by the declining utility of extra salary. Schaafsma *et al.* (2014) define the average transfer error of MWTP as the average percentage change in MWTP between waves. Our average transfer error is 59% between W2 and W1, and 11% between the counterfactual case and W1. The latter compares well against -35% in Schaafsma *et al.* (2014). Our own computation for other studies reporting MWTP finds 23% in Liebe *et al.* (2012), whose study rejected the temporal stability of utility weights as Schaafsma *et al.* did,

[9]This expression corresponds to $1 - \lambda$, where λ solves $\theta_{nk} + \theta_{nS}\ln(\lambda \times salary) = \theta_{nS}\ln(salary)$.

Figure 2: Mean willingness-to-pay (WTP)



The WTP refers to the fraction of salary that the decision maker is willing to trade off for an improvement in a particular attribute. Figures in the brackets, (A,B,C) indicate that the mean WTP in wave 1 is A; the mean WTP in the counterfactual case, that uses wave 2 non-salary coefficients and wave 1 salary coefficients, differs from A by B; and the actual mean WTP in wave 2 differs from A by C.

and -36% in Skjoldborg *et al.* (2009) and 4% in Czajkowski *et al.* (2014) whose studies did not.[10,11]

In sum, the average transfer error in MWTP for nursing job attributes would have been towards the lower end of what previous studies have found for other types of attributes, if not for the instability of preferences for salary. Before concluding, we provide further thoughts on this instability.

To ensure that the declining utility of extra salary is a feature of the underlying data, not that of our $ln(salary)$ specification, a robustness check is performed. We modify

[10]Skjoldborg *et al.* (2009) and Czajkowski *et al.* (2014) conduct 3 DCEs over their study periods. These results are based on their first and last DCEs.

[11]A modified version of the metric of Schaafsma *et al.* (2014) that averages the absolute values of percentage changes instead of percentage changes leads to similar results.

the LCL model above by replacing $ln(salary)$ with 6 dummy variables (3 for each wave) capturing all salary increments. The online appendix reports the results, which show that a given amount of extra salary in W2 yields either the same or less utility gain than a smaller amount in W1. The online appendix also reports specifications that incorporate observed heterogeneity in utility weights using individual characteristics. In general, the utility from extra salary and other attributes does not vary systematically across career stage groups (graduates and different years of study). This 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.[12]

One obvious and potentially testable confounding factor is the "price vector effects" (Hanley *et al.*, 2005), which arise when elicited preferences are sensitive to the levels of a monetary attribute used in the DCE design. As noted, four salary levels in our design have increased for every respondent between W1 and W2, to keep the choice scenarios plausible in relation to the updated pay schedule. Distinguishing the price vector effects from the temporal shift in preferences for salary requires randomly allocating respondents to two different salary vector treatments within the same wave; this would be a useful avenue to pursue in future work.

If present, sizable price vector effects would have implications beyond explaining the temporal instability of preferences for salary. Specifying a small number of salary levels is a common feature of DCEs in health workforce research (e.g. Blaauw *et al.* (2010), Kolstad (2011), and Sivey *et al.* (2012) consider four levels), and such effects would limit the external validity of the resulting WTP estimates. Of three previous studies that have tested for the price vector effects in the contexts of cervical cancer screening (Ryan and Wordsworth, 2000) and environmental valuation (Hanley *et al.* 2005; Carlsson and Martinsson, 2008), only one study (Carlsson and Martinsson, 2008) reports significant effects.

The findings of our earlier study (Yoo and Doiron, 2013) suggest that the robust elicitation of preferences for salary could be an inherently difficult task. In that study, we compare the DCE above against another type of DCE that presents respondents with

[12]Interestingly, first-year students, who spent the least time in practicum placements, are the only group that differs from others in that first-year students place more weights on the hospital's reputation for excellent quality of care.

only one nursing job profile at a time, asking them to state the best and worst aspects of that particular job. The same respondents complete both types of DCEs which use the same first-wave salary vector. Weights on non-salary attributes are comparable across the two types of DCEs but again, the salary weights differ. These results, along with the present analysis, prompt us to recommend further work on the stability of preferences for monetary attributes generally.

4 Conclusion

This paper is the first study on the temporal stability of stated job preferences, an important reliability criterion for the intended policy use of a growing number of DCE studies in health workforce research. Statistically, we find stable preference parameters for only 7 of 12 job attributes in our DCE. However, the relative importance of different attributes to job choices is stable enough to support the use of DCEs to identify priority areas of intervention. Our conclusion regarding the stability of willingness-to-pay (WTP) for non-salary attributes is more tentative. The average transfer error in our WTP estimates is larger than what temporal stability studies in other contexts have found. But this discrepancy is mostly explained by the instability of preferences for salary: in the absence of it, the average transfer error is 11%, well within the lower end of what other studies have found. We believe that a research agenda focusing on the difficulty of eliciting preferences for monetary characteristics may be warranted.

Acknowledgements

We wish to thank Editor James F. Burgess and two anonymous referees for helpful and constructive comments. We wish to thank Agne Suziedelyte for helpful suggestions. We also wish to thank participants at the iHEA World Congress in 2013.

References

Akaich 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.

Barnes AJ, Hanoch Y, Rice T. 2015. Determinants of coverage decisions in health insurance marketplaces: consumers decision-making abilities and the amount of information in their choice environment. *Health Services Research* **50**: 58-80.

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.

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

Caparrós A, Oviedo JL, Campos P. 2008. Would you choose your preferred option? comparing choice and recoded ranking experiments. *American Journal of Agricultural Economics* **90**: 843-855.

Carlsson F, Martinsson P. 2008. How much is too much? *Environmental and Resource Economics* **40**: 165-176.

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.

DeShazo JR, Fermo G. 2002. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and Management* **44**: 123-143.

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

Hanley N, Adamowicz W, Wright R. 2005. Price vector effects in choice experiments: an empirical test. *Resource and Energy Economics* **27**: 227-234.

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.

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.

Kelly J, Ahern K. 2009. Preparing nurses for practice: a phenomenological study of the new graduate in Australia. *Journal of Clinical Nursing* **18**: 910-918.

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.

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

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.

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, Wordsworth S. 2000. Sensitivity of willingness to pay estimates to the level of attributes in discrete choice experiments. *Scottish Journal of Political Economy* **47**: 504-524.

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.

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

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

Online Appendix for:

Temporal stability of stated preferences: the case of junior nursing jobs

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

a. School of Economics, University of New South Wales

Sydney, Australia. d.doiron@unsw.edu.au

b. Durham University Business School, Durham University

Durham, United Kingdom. h.i.yoo@durham.ac.uk

January 2016

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

1 Supporting material for the main manuscript

Table A.1 provides a full summary of the attributes and levels used in the DCE design. Table A.2 reports detailed estimation results for our main latent class logit (LCL) model where weekly salary enters as $\ln(salary)$; Figure 1, Figure 2 and Table 1 in the main manuscript are derived from these results.

2 Additional results: other mixed logit models

Our analytic framework is the usual random utility maximisation model. The utility that person n derives from job j in choice scenario t is specified as

$$\begin{aligned} U_{njt} &= \boldsymbol{\beta}'_{nt}\boldsymbol{x}_{njt} + \varepsilon_{njt} \\ \text{where } \boldsymbol{\beta}_{nt} &= \boldsymbol{\alpha}_n + \boldsymbol{\delta}_n \times wave2_{nt.} \end{aligned} \tag{1}$$

$\boldsymbol{x}_{njt}$ is a vector of 12 attributes, $\boldsymbol{\beta}_{nt}$ is a conformable vector of utility coefficients, $wave2_{nt}$ is a dummy variable that equals 1 for choice scenarios from wave 2, and error term ε_{njt} is i.i.d. Type 1 Extreme Value. $\boldsymbol{\beta}_{nt}$ is further decomposed into $\boldsymbol{\alpha}_n$ that captures the baseline preferences in wave 1, and $\boldsymbol{\delta}_n$ that captures the deviations from the baseline in wave 2.

The main manuscript focuses on the LCL model that uses a discrete distribution to approximate the joint distribution of $\boldsymbol{\alpha}_n$ and $\boldsymbol{\delta}_n$ nonparametrically.[1] In this section, we present additional results based on other types of mixed logit models that place more parametric restrictions on the joint distribution. Specifically, we estimate the normal-mixture logit (NMIXL) model that specify $\boldsymbol{\beta}_{nt}$ as

$$\boldsymbol{\beta}_{nt} = \overline{\boldsymbol{\alpha}} + \overline{\boldsymbol{\delta}} \times wave2_{nt} + \boldsymbol{\mu}_n \tag{2}$$

where $\overline{\boldsymbol{\alpha}}$ is a vector of mean utility coefficients in wave 1; $\overline{\boldsymbol{\alpha}} + \overline{\boldsymbol{\delta}}$ is a vector of mean utility coefficients in wave 2; and $\boldsymbol{\mu}_n$ is a conformable vector of draws from a zero-mean multivariate normal distribution, $MVN(\mathbf{0}, \boldsymbol{V})$, that captures individual-specific deviations from the population mean coefficients.[2] The k^{th} element of $\boldsymbol{\mu}_n$, μ_{kn}, then measures by how much person n's utility coefficient on the k^{th} attribute deviates from the mean

[1] The estimation results have been obtained using Stata command -lclogit- (Pacifico and Yoo, 2013).
[2] The estimation results have been obtained using Stata command -mixlogit- (Hole, 2007).

coefficient on that attribute, and the l^{th} element μ_{ln} can be interpreted similarly. The k^{th} diagonal element of $\boldsymbol{V}$ is the population variance of μ_{kn}, and indicates how dispersed the taste for the k^{th} attribute (as measured by the utility derived from that attribute) is across individuals. The off-diagonal element in the k^{th} row and the l^{th} column of $\boldsymbol{V}$ is identical to the element in the l^{th} row and the k^{th} column, as they correspond to the population covariance of μ_{kn} and μ_{ln}. A positive (negative) covariance means that someone who derives a large utility from the k^{th} attribute tends to derive a large (small) utility from the l^{th} attribute; in this sense, a non-zero covariance indicates correlated tastes for two underlying attributes.

We also estimate the Generalized Multinomial Logit (GMNL) model of Fiebig *et al.* (2010) that augments NMIXL with an additional random parameter σ_n to accommodate interpersonal heterogeneity in the overall scale of utility. This model specify $\boldsymbol{\beta}_{nt}$ as

$$\boldsymbol{\beta}_{nt} = \sigma_n(\overline{\boldsymbol{\alpha}} + \overline{\boldsymbol{\delta}} \times wave2_{nt} + \boldsymbol{\mu}_n) \tag{3}$$

where $\ln(\sigma_n)$ is a draw from a univariate normal distribution $N(-0.5\tau^2/, \tau^2)$, and other notations are the same as in the NMIXL context.[3]

The left panel of Figure A.1 plots the mean coefficients from an NMIXL specification that assumes away correlated tastes for different attributes. This specification has $\overline{\boldsymbol{\alpha}}, \overline{\boldsymbol{\delta}}$, and the diagonal elements of $\boldsymbol{V}$ as parameters to estimate, and constrains each off-diagonal element of $\boldsymbol{V}$ to 0. The right panel of Figure A.1 plots the mean coefficients from a GMNL specification which extends the preceding NMIXL specification by including τ as an additional parameter to estimate.

Similarly, the left panel of Figure A.2 plots the mean coefficients from an NMIXL specification that allows for correlated tastes. This specification has $\overline{\boldsymbol{\alpha}}, \overline{\boldsymbol{\delta}}$, the diagonal elements of $\boldsymbol{V}$, and all distinct off-diagonal elements of $\boldsymbol{V}$ as parameters to estimate. The right panel of Figure A.2 plots the mean coefficients from a GMNL specification which extends this more general NMIXL specification by including τ as an additional parameter to estimate.

In general, both figures look similar to Figure 1 in the main manuscript and suggest that qualitatively the same pattern of temporal variations in the mean coefficients can be detected regardless of which mixed logit model is used. While there are some differences in relation to which specific coefficients show statistically significant variations, they do not affect our main conclusion that a majority of the coefficients do not show significant

[3] The estimation results have been obtained using Stata command -gmnl- (Gu *et al.*, 2013).

variations; that the relative importance of different attributes remains fairly stable; and that there is a substantial decline in the utility from extra salary which deserves further research.

3 Additional results: dummy salary specification

As discussed in the main manuscript, the average transfer error in the mean willingness-to-pay (MWTP) for nursing job attributes would have been towards the lower end of what previous studies have found for other types of attributes, if not for the instability of preferences for salary. To ensure that the declining utility of extra salary is a feature of the underlying data, not that of our $ln(salary)$ specification, we perform a robustness check. Specifically, we modify our 3-class LCL model by replacing $ln(salary)$ with dummies for all salary increment levels in each wave, thereby making the model specification non-parametric with respect to salary gains. Figure A.3 plots the mean utility weights from the resulting 3-class latent class logit (LCL) model. The weights on salary gains are not directly comparable between waves, as salary changes by an increment of \$150 in W1, and \$200 in W2. Nevertheless, Figure 3 clearly illustrates that a given amount of extra salary in W2 yields either the same (on the 45° line) or less (below the 45° line) utility gains than a smaller amount in W1.

4 Additional results: demographic specifications

Our two DCE waves were spaced at least a full year apart, and 15 months on average. By interacting each attribute with personal characteristics and using those interaction terms as additional regressors, we can allow the utility coefficients to vary with personal characteristics that vary between waves (e.g. stage of career), as well as time-invariant (e.g. gender) characteristics. In case preferences vary systematically with time-varying characteristics, it may be possible to explain some temporal shifts in preferences with reference to changes in those characteristics between two waves.

Note, however, that since there are 12 attributes, incorporating one characteristic requires estimating at least 12 extra parameters. Unless further restrictions are placed on the selection of the demographic interaction terms, incorporating even a moderately large number of characteristics can make the resulting specification susceptible to overfitting and difficult to interpret and present.

To achieve a parsimonious demographic specification, we followed a general-to-specific specification search strategy. As it is computationally impractical to estimate a mixed logit model repeatedly, our specification search was based on the multinomial logit (MNL) model.[4] We started off with a general MNL specification that involved 300 regressors: 12 attributes, 12 interaction terms between attributes and the wave 2 dummy, and 252 interaction terms involving attributes and 21 demographic variables. These 21 demographic variables incorporate both time-varying and time-invariant personal characteristics.[5] The general specification was estimated twice, once using the balanced estimation sample of 241 respondents who completed the DCE in both waves and provide 3821 choice observations, and once using the unbalanced sample that included additional 387 respondents who only completed the first-wave DCE and provide extra 3096 choice observations.[6] Based on the results, we obtained a simplified specification that retained all 12 attributes and those out of 264 interaction terms which were significant at the 10% level in either estimation sample. We then proceeded to testing down the MNL model repeatedly in an analogous manner, by using a simplified specification from one stage as a new general specification for the next stage, until we obtained a specification wherein all retained interaction terms became significant at the 10% level in either estimation sample.

The resulting final MNL specification involves 28 regressors: 12 attributes, 4 interaction terms between attributes (ln($salary$), "well staffed", "well equipped", "abundant parking") and the wave 2 dummy, and 12 demographic interaction terms. An earlier version of this paper reports the MNL estimates based on this specification for both balanced and unbalanced samples.[7]

The current main manuscript focuses on the analysis of temporal stability in preferences using the balanced sample. Accordingly, we modify the above final specification

[4]The estimation run time for our mixed logit models varied from 4 hours for the LCL model to 35 hours for the Correlated GMNL model in Figure A.2.

[5]The demographic variables covered: (1-3) career stage, captured by three dummies; (4) university affiliation at the time of survey recruitment; (5) gender; (6-10) five dummies related to the respondent's current and planned parental status; (11) self-assessed health; (12) marital status; (13-14) two partner characteristics (whether living with the respondent, whether employed); (15) satisfaction with current nursing job if the respondent has one; (16-17) income and missing income flag; (18) whether born in Australia; (19) whether international student; (20) whether English is native language; (21-24) four age group dummies. More information is available upon request.

[6]The unbalanced sample, with more people, offers a greater scope for detecting observed heterogeneity of interest in this specification search process. Our main manuscript focuses on the balanced sample, however, for ease of interpretation of the results as temporal shifts in preferences.

[7]The earlier version can be accessed at: `https://goo.gl/bAkumy`.

further as follows. First, we drop 4 demographic interactions terms which are significant in the unbalanced sample, but insignificant in the balanced sample.[8] Second, we include all 12 interaction terms between attributes and the wave 2 dummy, to facilitate more direct comparisons with our baseline models that do not place any prior restriction on which utility coefficient varies over time. The new MNL specification involves 32 regressors: 12 attributes, 12 interaction terms between attributes and the wave 2 dummy, and 8 demographic interaction terms.

Table A.3 reports the coefficient estimates from this last MNL specification, and the mean coefficient estimates from its 3-class LCL model extension. Figure A.4 plots the wave 2 estimates from each model against the corresponding wave 1 estimates, for the reference group of respondents who are female, aged 22-29, not in first-year of university, and have no children under 6. In general, conditioning on observed taste heterogeneity through demographic interaction terms does not affect our baseline findings regarding temporal stability, as all previously significant temporal variations except one remains significant.[9]

That only a small number of demographic interaction terms turn out to be significant is not uncommon in DCE studies, both in the context of health workforce research (Sivey *et al.*, 2012; Holte *et al.*, 2015) and other contexts (Liebe *et al.*, 2012; Schaafsma *et al.*, 2014). In the present context, however, the general lack of career stage (first-year, second-year, third-year students and graduate nurses) effects may appear surprising if one takes a view that job preferences would vary with job experiences, though our result agrees with the findings of two previous temporal stability studies (San Miguel *et al.*, 2002; Liebe *et al.*, 2012) which tested for the effects of relevant experiences in their contexts and found no significant effects.[10] We note that when 35 omitted interaction terms involving attributes and career stage dummies are added to the MNL specification

[8] They are interaction terms involving: (1) "excellent quality of care" and "studied at the University of New England" (base group: "studied at the University of Technology Sydney"); (2) "flexible rostering" and "first-year students"; (3) "flexible rostering" and "male"; (4) "3 rotations" and "first-year students".

[9] In the baseline MNL model that excludes demographic heterogeneity, we found a significant change in the utility weight on "excellent quality of care" that does not show up in Table A.3 and Figure A.4. In the LCL model that excludes demographic heterogeneity (Figure 1 in the main manuscript), the change in the mean utility weight on "supportive management" was significant at the 5% level; it is now significant only at the 10% level.

[10] The alternatives in San Miguel *et al.* (2002) describe out-of-hours GP services, and the relevant experience is whether the respondent used such a service between waves. The alternatives in Liebe *et al.* (2012) describe landscape outcomes associated with wind power generation, and the relevant experience is whether the respondent was exposed to wind power generation issues between waves.

in Table A.3 as additional regressors, the log-likelihood of the MNL model increases only slightly to -3036.14 from -3050.96: the LR statistic of 29.64 has a $\chi^2(35)$ p-value of 0.72, and does not reject our current specification at any conventional significance level. We offer two thoughts on the general lack of career stage effects. First, this may suggest a good understanding of nursing jobs by nursing students acquired through the practicum component of the Bachelor of Nursing program. 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. Second, this may suggest that both experiences and/or preference changes with experiences are individual-specific, instead of career stage-specific. Note that the career stage interaction terms are only able to pick up systematic variations in preferences across career stage groups.

References

Fiebig DG, Keane M, Louviere J, Wasi N. 2010. The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science* **29**: 393-421.

Gu Y, Hole AR, Knox S. 2013. Fitting the generalized multinomial logit model in Stata. *Stata Journal* **13**:382-397.

Hole AR. 2007. Fitting mixed logit models by using maximum simulated likelihood. *Stata Journal* **7**: 388-401.

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.

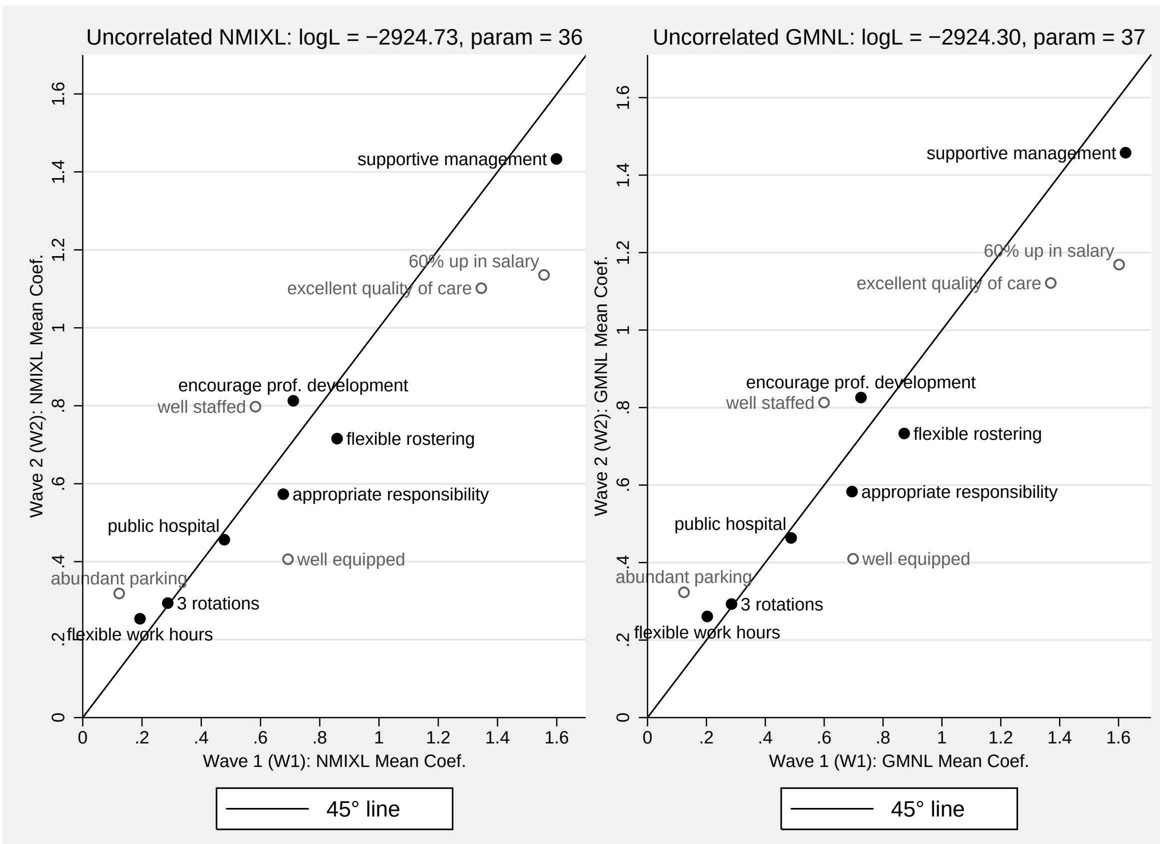
Liebe U, Meyerhoff J, Hartje V. 2012. Test-retest reliability of choice experiments in environmental valuation. *Environmental and Resource Economics* **53**: 389-407.

Pacifico D, Yoo HI. 2013. lclogit: A Stata command for fitting latent-class conditional logit models via the expectation-maximization algorithm. *Stata Journal* **13**: 625-639.

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

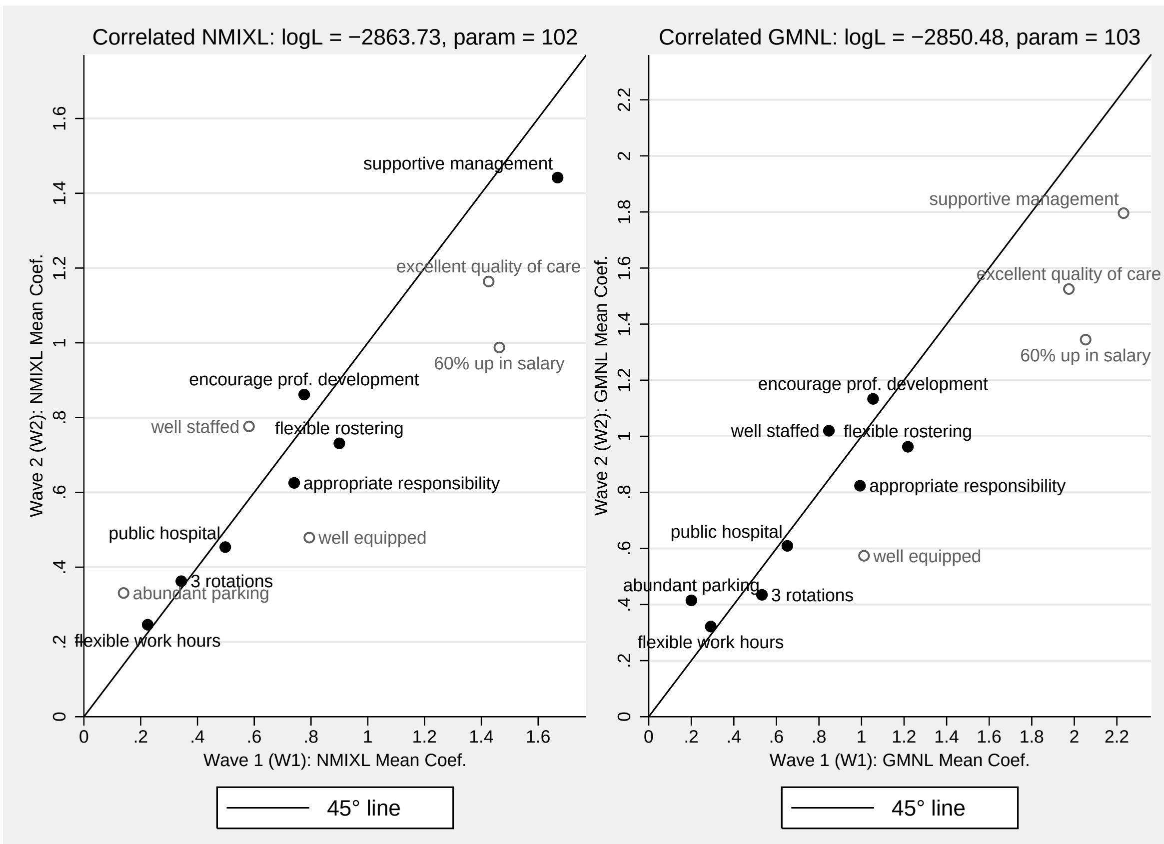
Sivey P, Scott A, Witt J, Joyce C, Humphreys J. 2012. Junior doctors' preferences for specialty choice. *Journal of Health Economics* **31**: 813-823.

Figure A.1: Uncorrelated mixed logit: wave-specific mean coefficients



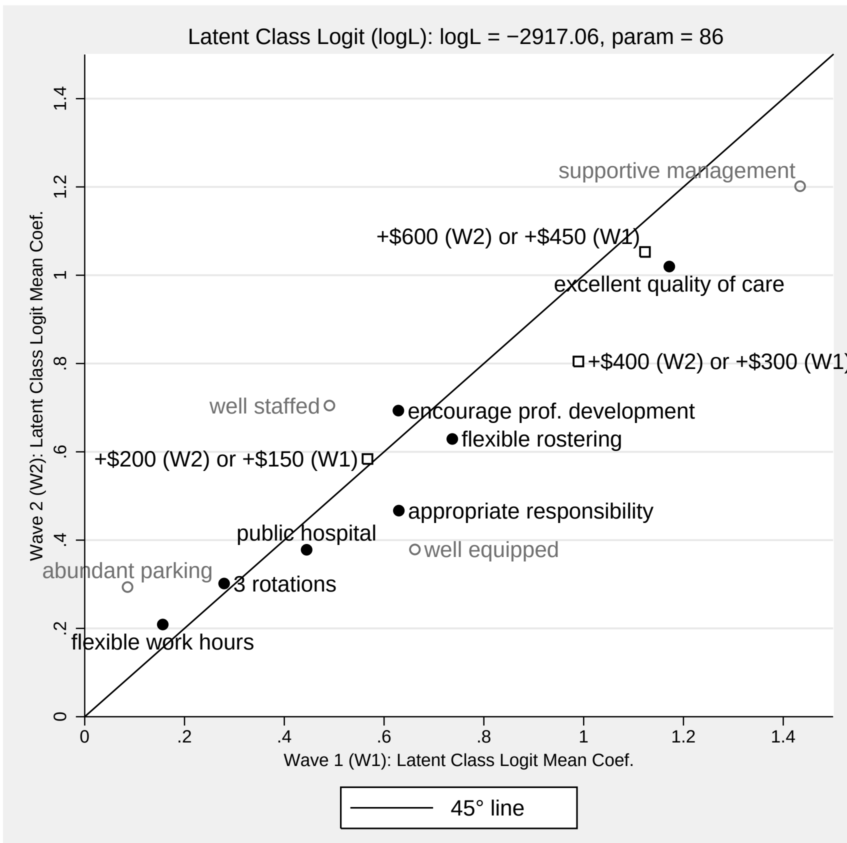
The sample includes 3821 choice observations from 241 individuals. LogL (param) refers to the maximised log-likelihood (number of estimated parameters). The log-likelihood has been simulated using 1000 random draws. A hollow circle indicates a significant between-wave difference at the 5% level. Detailed estimation results are available upon request.

Figure A.2: Correlated mixed logit: wave-specific mean coefficients



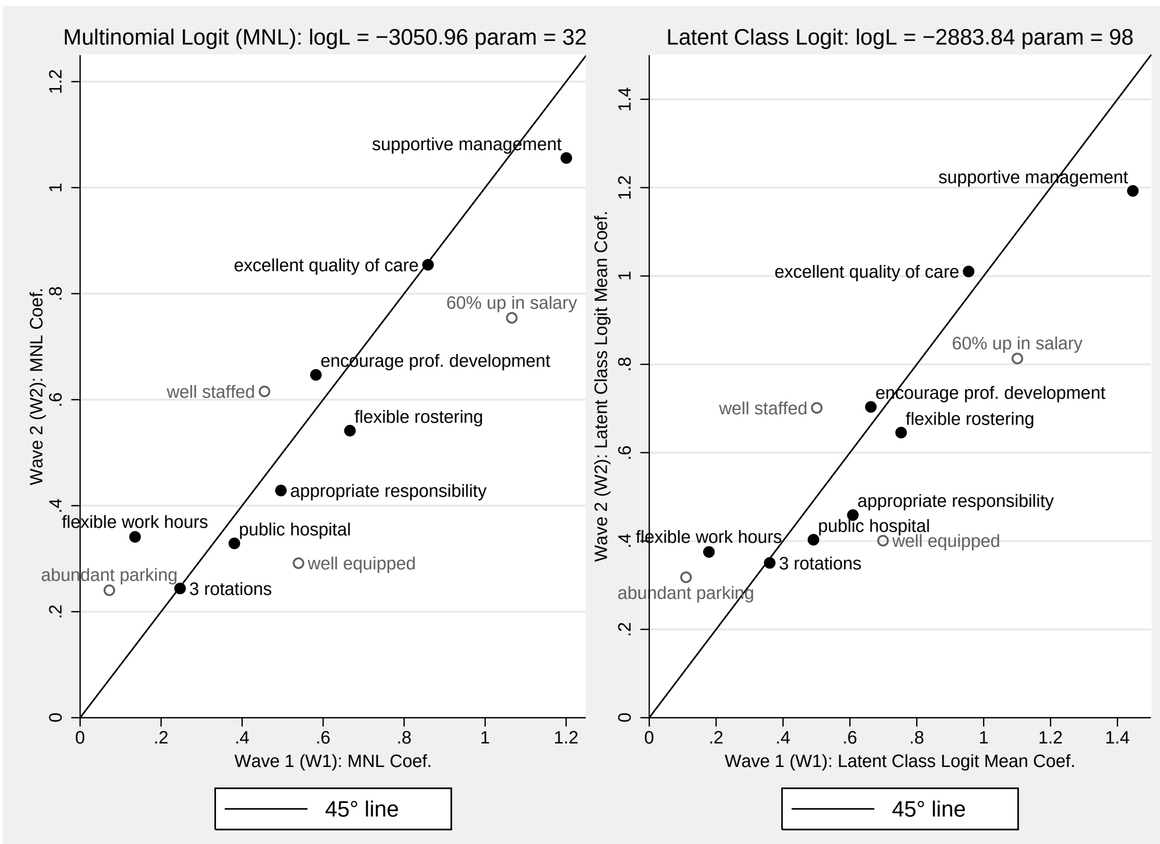
The sample includes 3821 choice observations from 241 individuals. LogL (param) refers to the maximised log-likelihood (number of estimated parameters). The log-likelihood has been simulated using 1000 random draws. A hollow circle indicates a significant between-wave difference at the 5% level. Detailed estimation results are available upon request.

Figure A.3: Wave-specific mean coefficients when specifying salary-level dummies



The sample includes 3821 choice observations from 241 individuals. Based on a 3-class latent class logit model specification that replaces $ln(salary)$ with three salary dummies per each wave. A hollow circle (diamond) indicates a significant between-wave difference at the 5% level. A hollow square indicates that the estimates are not comparable between waves. Detailed estimation results are available upon request.

Figure A.4: Demographic specifications: wave-specific (mean) coefficients



The sample includes 3821 choice observations from 241 individuals. LogL (param) refers to the maximised log-likelihood (number of estimated parameters). A hollow circle indicates a significant between-wave difference at the 5% level. The results are for the reference group of respondents who are female, aged 22-29, not in first-year of university, and have no children under 6. See Table A.3 for further results.

Table A.1: Job attributes and associated levels

Glossary definition of attribute	**Attribute name**	**Levels**
The type of hospital where the new graduate program is located	Location	Private hospital Public hospital
The number of rotations to different clinical areas	Clinical rotations	None Three
Whether the new graduate program offers fulltime and part-time positions, or fulltime only	Work hours	Fulltime only Part-time or fulltime
The flexibility of the rostering system in accommodating requests	Rostering	Inflexible, does not allow requests Flexible, usually accommodating requests
The hospital's reputation regarding staffing levels	Staffing levels	Frequently short of staff Usually well-staffed
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
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
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
The parking facilities	Parking	Limited Abundant and safe
The hospital's reputation regarding the responsibility given to nurses, relative to their qualifications and experience	Responsibility	Too much responsibility Appropriate responsibility
The hospital's reputation regarding the quality of patient care	Quality of care	Poor Excellent
The gross weekly salary	Salary	Wave 1: \$800; \$950; \$1100; \$1250 Wave 2: \$900; \$1100; \$1300; \$1500

Table A.2: Latent class logit: main specification

	Class 1		Class 2		Class 3		Mean	
ln(salary)	4.477***	(0.465)	1.398***	(0.440)	1.839***	(0.467)	2.555***	(0.241)
int: *wave2	-0.679	(0.553)	0.406	(0.571)	-1.496**	(0.593)	-0.611**	(0.302)
supportive management	0.614***	(0.121)	2.418***	(0.294)	1.185***	(0.167)	1.400***	(0.103)
int: *wave2	-0.196	(0.168)	-0.336	(0.322)	-0.169	(0.210)	-0.232**	(0.117)
excellent quality of care	0.612***	(0.125)	0.747***	(0.143)	2.046***	(0.197)	1.156***	(0.083)
int: *wave2	-0.062	(0.174)	-0.252	(0.193)	-0.177	(0.258)	-0.164	(0.100)
flexible rostering	0.927***	(0.133)	0.757***	(0.133)	0.511***	(0.143)	0.727***	(0.067)
int: *wave2	-0.259	(0.184)	-0.251	(0.177)	0.186	(0.188)	-0.101	(0.094)
encourage prof. dev.	0.524***	(0.115)	0.597***	(0.130)	0.797***	(0.145)	0.643***	(0.068)
int: *wave2	-0.235	(0.161)	0.378**	(0.189)	-0.014	(0.189)	0.042	(0.097)
approp. responsibility	0.153	(0.125)	0.891***	(0.172)	0.772***	(0.141)	0.609***	(0.077)
int: *wave2	0.370**	(0.174)	-0.503**	(0.219)	-0.306	(0.190)	-0.150	(0.103)
well staffed	0.523***	(0.114)	0.514***	(0.137)	0.444***	(0.133)	0.492***	(0.065)
int: *wave2	0.239	(0.162)	0.182	(0.183)	0.170	(0.179)	0.196**	(0.092)
well equipped	0.694***	(0.126)	0.672***	(0.218)	0.653***	(0.148)	0.673***	(0.081)
int: *wave2	-0.314*	(0.180)	-0.385	(0.252)	-0.174	(0.194)	-0.288***	(0.107)
public hospital	0.158	(0.121)	0.736***	(0.208)	0.487***	(0.134)	0.461***	(0.079)
int: *wave2	0.081	(0.164)	-0.540**	(0.267)	0.205	(0.179)	-0.078	(0.107)
3 rotations	0.205*	(0.121)	0.550***	(0.192)	0.189	(0.134)	0.312***	(0.079)
int: *wave2	0.117	(0.173)	0.001	(0.255)	-0.114	(0.188)	-0.001	(0.106)
flexible work hours	0.197*	(0.115)	0.015	(0.116)	0.335**	(0.133)	0.186***	(0.064)
int: *wave2	0.242	(0.166)	-0.051	(0.165)	-0.061	(0.182)	0.041	(0.089)
abundant parking	0.170	(0.114)	0.001	(0.120)	0.132	(0.124)	0.102	(0.063)
int: *wave2	0.307*	(0.165)	0.154	(0.168)	0.154	(0.172)	0.204**	(0.089)
Class Share	0.326***	(0.042)	0.325***	(0.043)	0.349***	(0.044)		
log-likelihood	-2927.24							
estimated parameters	74							

Standard errors in parenthesis. *, **, *** indicate statistical significance at the 10%, 5% and 1% levels. int: is the interaction of the wave 2 dummy and the preceding job attribute. The model estimates the coefficients on the job attributes and the interaction terms for Class 1, Class 2, and Class 3 (72 parameters in total); and the class shares for Class 1 and Class 2 (2 parameters in total). Other results are derived from these 74 parameters. The sample includes 3821 choice observations from 241 individuals.

Table A.3: Demographic specifications

	MNL		LCL mean	
ln(salary)	2.268***	(0.239)	2.342***	(0.252)
int: *wave2	-0.663***	(0.248)	-0.612**	(0.306)
int: *male	1.098**	(0.503)	1.169**	(0.489)
int: *born o/s	1.694***	(0.540)	1.552***	(0.557)
supportive management	1.201***	(0.080)	1.446***	(0.106)
int: *wave2	-0.145	(0.093)	-0.253**	(0.119)
excellent quality of care	0.859***	(0.086)	0.955***	(0.092)
int: *wave2	-0.004	(0.104)	0.055	(0.112)
int: *first-year	0.488***	(0.143)	0.571***	(0.145)
flexible rostering	0.666***	(0.069)	0.753***	(0.068)
int: *wave2	-0.125	(0.086)	-0.108	(0.095)
encourage prof. development	0.582***	(0.069)	0.663***	(0.069)
int: *wave2	0.065	(0.086)	0.041	(0.098)
appropriate responsibility	0.496***	(0.068)	0.609***	(0.078)
int: *wave2	-0.067	(0.088)	-0.150	(0.104)
well staffed	0.455***	(0.054)	0.501***	(0.065)
int: *wave2	0.160**	(0.074)	0.200**	(0.093)
well equipped	0.539***	(0.061)	0.699***	(0.083)
int: *wave2	-0.248***	(0.082)	-0.298***	(0.109)
public hospital	0.381***	(0.067)	0.491***	(0.079)
int: *wave2	-0.052	(0.088)	-0.088	(0.106)
3 rotations	0.247***	(0.061)	0.360***	(0.082)
int: *wave2	-0.003	(0.079)	-0.010	(0.108)
int: *born o/s	-0.372***	(0.135)	-0.457**	(0.192)
flexible work hours	0.136*	(0.081)	0.179*	(0.096)
int: *wave2	0.017	(0.074)	0.018	(0.092)
int: *aged $\leq$ 21	-0.189**	(0.091)	-0.179*	(0.108)
int: *aged 30-39	0.072	(0.123)	0.081	(0.133)
int: *age $\geq$ 40	0.127	(0.153)	0.076	(0.153)
int: *has child under 6	0.581***	(0.162)	0.706***	(0.170)
abundant parking	0.072	(0.052)	0.110*	(0.064)
int: *wave2	0.168**	(0.077)	0.208**	(0.090)
log-likelihood	-3050.96		-2883.84	
parameters	32		98	

Standard errors in parenthesis. *, **, *** indicate statistical significance at the 10%, 5% and 1% levels. Int: is the interaction of the namesake dummy regressor and the preceding job attribute. The base person is a female first-wave respondent aged 22-29, is not a first-year student, does not care for a child under 6, and was born in or migrated to Australia 3 years before her enrollment at the U. of Technology Sydney. MNL reports the coefficient estimates from the multinomial logit model. LCL mean reports the mean coefficient estimates from the 3-class latent class logit model; detailed estimation results are available upon request. The sample includes 3821 choice observations from 241 individuals.