Dimitris Potoglou, Peter Burge, Terry Flynn, Ann Netten, Juliette Malley, Julien Forder and John E. Brazier

Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data

Article (Accepted version)
(Refereed)

Original citation:
Potoglou, Dimitris and Burge, Peter and Flynn, Terry and Netten, Ann and Malley, Juliette and Forder, Julien and Brazier, John E. (2011) *Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data.* *Social science & medicine*, 72 (10). pp. 1717-1727. ISSN 0277-9536
DOI: 10.1016/j.socscimed.2011.03.027

Best-Worst Scaling vs. Discrete Choice Experiments: An Empirical Comparison using Social Care Data

Authors: Dimitris Potoglou[1], Peter Burge[1], Terry Flynn[2], Ann Netten[3], Juliette Malley[3], Julien Forder[3], John E Brazier[4],

Author affiliations:

[1] RAND Europe, Cambirdge

[2] Centre for the Study of Choice, University of Technology Sydney

[3] Personal Social Services Research Unit, University of Kent

[4] Heath Economics and Decision Science, School of Health and Related Research, University of Sheffield

Corresponding author:

Dimitris Potoglou

Accepted for publication by Social Science and Medicine

Key messages

- This study illustrates key issues that are important in choosing between profile-case best-worst scaling and discrete choice experiment studies
- Empirical research on the value of outcomes of social care reveals similar patterns in the preference weights obtained from the two approaches
- In the majority of cases examined, preference weights are not significantly different once the weights have been appropriately normalised/rescaled

Abstract

This paper presents empirical findings from the comparison between two principal preference elicitation techniques: discrete choice experiments and profile-based best-worst scaling. Best-worst scaling involves less cognitive burden for respondents and provides more information than traditional "pick-one" tasks asked in discrete choice experiments. However, there is lack of empirical evidence on how best-worst scaling compares to discrete choice experiments. This empirical comparison between discrete choice experiments and best-worst scaling was undertaken as part of the Outcomes of Social Care for Adults project, which aims to develop a weighted measure of social care outcomes. The findings show that preference weights from best-worst scaling and discrete choice experiments do reveal similar patterns in preferences and in the majority of cases preference weights - when normalised/rescaled - are not significantly different.

Keywords
UK; Best-worst scaling; discrete choice experiments; stated choice; discrete choice models; social care, social care outcomes; quality of life

Introduction

Priority-setting in many areas of public policy is informed through the use of public preferences. Within the 'non-welfarist' or 'extra-welfarist' paradigm, public preferences are elicited in nationally representative valuation exercises, typically using the standard gamble (SG) or time trade-off (TTO) (Brazier, Ratcliffe, Salomon, & Tsuchiya, 2007). These tools require respondents to manipulate probabilities or lengths of life and so rely on an assumption of cardinality in responses. Theoretical and empirical problems with these methods (Bleichrodt, 2002) have led to interest in tasks that require only ordinality in responses, such as discrete choice experiments (DCEs) and ranking studies. DCEs have been used extensively to facilitate analyses in the fields of transport and environmental policy. However, they can also be used to value different instruments and work is underway to do so for the EQ-5D-5L for measuring health outcomes (as a supplement to a TTO valuation) and the ICECAP capability indices (at least).

Best-worst scaling (BWS) is an alternative preference elicitation method that also only requires an assumption of ordinality. It was developed by Louviere and Woodworth (1990) and its first application was published in 1992 (Finn & Louviere) illustrating Case 1 (the 'object' case). The method gained popularity in health and social care when the properties of Case 2 (the 'profile' – previously called 'attribute' – case) were proved (Marley, Flynn, & Louviere, 2008) and a guide to its use was published (Flynn, Louviere, Peters, & Coast, 2007). Flynn (2010a) provides an overview and theoretical discussion of the different cases of BWS. Case 2 has particular advantages in valuation studies that seek to elicit general population preferences for important attributes of quality of life (or whatever maximand is of relevance to policymakers). In particular, it presents profiles one at a time, rather than in choice sets of size two or more as in a traditional DCE. This is important when respondents do not have experience of making choices in the particular area of application: keeping two or more profiles in mind at once is likely to be a harder task, leading to an increase the size

of the random utility component and reduction of the statistical efficiency of the preference elicitation.

This paper reports on the empirical comparison between the discrete choice and profile-case BWS experiments using data from a pilot study seeking to elicit values for different dimensions of social care related quality of life. The specific objective is to determine the extent to which valuations of quality of life states obtained through a best-worst scaling experiment are comparable to those obtained through discrete choice experiments. To our knowledge, this paper is the first to empirically test the comparability of the profile-case BWS and DCE estimates. The following section provides a brief background to general research framework. Then the methods employed in this study including the design of the discrete choice and best-worst scaling experiments, data collection and econometric analysis will be presented. Finally, model estimations from the discrete choice and best-worst experiments and the comparison of values between DCE and BWS will be discussed.

Background to ASCOT measure

This research is part of the Outcomes of Social Care for Adults (OSCA) project (Netten, Malley, Forder, Burge, Potoglou, Brazier et al., 2009), which is building on work that has been undertaken on social-care outcome measurement over a number of years, including, the Individual Budget pilot evaluation (Glendinning, Challis, Fernández, Jacobs, Jones, Manthorpe et al., 2008). The measure being developed is part of the Adult Social Care Outcome Toolkit (ASCOT) (see Netten et al., 2009 for full details). The toolkit is being developed as part of the Measuring Outcomes for Public Service Users (MOPSU) project, which was led by the Office for National Statistics (ONS) in UK (ONS, 2010). The work on adult social care focuses primarily on outcomes for residents of care homes (Netten, Beadle-Brown, Trukeschitz, Towers, Welch, Forder et al., 2010) and low level interventions, that is low cost services usually targeted at people with low level needs, for example many day centres for older people (Caiels, 2010).

The ASCOT measure is designed to capture information about an individual's social care-related quality of life (SCRQOL). The aim is for the measure to be applicable across as wide a range of user groups and care and support settings as possible. In identifying and defining the domains the aim was to ensure the measure is sensitive to outcomes of social care activities. Evidence from consultation with service users, experts and policy-makers, as well as focus group work and interviews with service users indicated that the measure captures aspects of SCRQOL that are valued by service users (and policy-makers) (Bamford, Qureshi, Nicholas, & Vernon, 1999; Malley, Sandhu, & Netten, 2006; Miller, Cooper, Cook, & Petch, 2008; Netten, McDaid, Fernández, Forder, Knapp, Matosevic et al., 2005; Netten, Ryan, Smith, Skatun, Healey, Knapp et al., 2002; Qureshi, Patmore, Nicholas, & Bamford, 1998).

Methods

Social care domains and levels

Evidence from previous analyses (Bamford et al., 1999; Malley et al., 2006; Miller et al., 2008; Netten et al., 2005; Netten et al., 2002; Qureshi et al., 1998), conceptual development and results of the consultation with stakeholders and service users and carers (Netten et al., 2009) fed into selection of domains and their levels. Nine domains were finally selected to describe social-care related quality-of-life situations: food and drink, personal cleanliness, accommodation, safety, social participation, occupation, control, dignity, and living in own home (see Table 1).

[Table 1, about here]

Each domain had four (4) levels; except for *living in own home* that had only two levels (living in own home; not living in own home). For purposes of clarity and in order to avoid wording that may lead to some domains dominating the choices, the *dignity* domain was worded as

the way I am helped and the *employment and occupation* domain was presented as *use of my time.*

Discrete choice experiments

Asking respondents to trade-off across nine different quality-of-life domains in a single discrete choice experiment would be particularly challenging. One option would be to concentrate on a subset of domains in the experiment. However, this option would narrow the findings of the study, as would combining the domains. Moreover, it would not be feasible to compare the findings with those in the BWS experiment. We therefore decided to split the nine domains across two discrete choice experiments (DCE1 and DCE2) with overlap in some domains to allow models from the two experiments to be based on a common utility scale. The decision to split the domains was guided by findings of a previous study, which showed that this strategy produced consistent values and the utility parameters of overlapping domains were equal (Burge, Netten, & Gallo, 2010).

It was acknowledged that by splitting the attributes across two experiments, it would not be possible to examine interactions in preferences for domains in different experiments. Therefore, it was decided to make prior assumptions about which interactions were most likely to be significant when grouping the domains to allow scope for estimating these should they prove to be important. In addition, some pragmatic considerations were also considered in grouping the attributes; for example, safety, personal care and food and drink represent the core outcomes of social services, so an argument existed for grouping these together. Cleaning the house, social participation and being active/occupied was seen as at a less fundamental level in terms such as Maslov's hierarchy of needs (Grewal, Lewis, Flynn, Brown, Bond, & Coast, 2006), so an additional argument existed for grouping these together. The final allocation of the domains between the two choice experiments is shown in Table 2.

[Table 2, about here]

Both DCE 1 and DCE 2 used a forced-choice design. Choice pairs in each of the discrete choice experiments were specified using a D-efficient choice design developed in the SAS software[1] (Kuhfeld, 2009). The main motivation for choosing an efficient design over an orthogonal design was to minimise the expected standard errors in the choice models that utilised the data from the experiment (Puckett & Rose, 2009). While there is nothing inherently wrong with orthogonal designs, these require larger sample sizes to yield parameter estimates with comparable levels of significance in choice models (Bliemer & Rose, 2005). The SAS routine for specifying the fractional factorial design matrix started with the full factorial design matrix and used the modified Fedorov algorithm to optimise the expected variance matrix on the basis of a set of prior parameters (Cook & Natchtsheim, 1980; Fedorov, 1972). At this stage, it was assumed that all attribute-level prior-parameters were zero; however, for the next stage of the study we could consider using the estimated coefficients from this pilot to feed in as priors for this optimisation.

Each of the design matrices in DCE1 and DCE2 included 128 situations. These design matrices covered sufficient domain-level combinations to allow all two-way interactions between domain-levels in the same experiment to be estimated. Obviously, 128 situations for each experiment would be too difficult for one respondent to evaluate. Therefore the design matrices were divided into 16 blocks so that each respondent was presented with 8 situations per choice experiment. Blocking was performed using the SAS Block procedure, which alters the order of situations until no (canonical) correlations exist between the block number and domain-levels. Figure 1 shows examples of DCE1 and DCE2 social care situations.

[Figure 1, about here]

[1] Another option would be to use Burgess' (2007) optimal design procedure. However the existing online (free) version does not allow the specification of conditional rules that are essential to our design, e.g. avoid scenarios where all attribute-levels take the value of 1 or 4 or avoid duplicates. (see, Burgess, L. (2007). Discrete Choice Experiments [computer software]. Sydney: Department of Mathematical Sciences, University of Technology).

Best-worst scaling experiment

The best-worst scaling experiment contained the same attributes as the discrete choice experiments, but rather than splitting them in two groups, all nine domains were presented in a single situation. Choice situations in the BWS experiment were specified using an orthogonal main effects plan (OMEP) that allowed eight domains with four levels and the *living in own home* domain with two levels to be tested. The full plan consisted of 4^8x2^1 combinations and using an orthogonal design these were reduced to 32 situations. Similarly to the DCEs, the 32 situations were blocked, so that each respondent repeated the choice task for 12 social-care situations. Figure 2 shows an example of a BWS exercise, where the respondent was asked to choose the best and then the worst domain.

[Figure 2, about here]

At this stage, it was important to acknowledge the potential for bias resulting from differences in the size of the random utility component variance by attribute. For example, presenting a situation in which one attribute was at its 'top' level, whilst all other attributes were at an intermediate level, was likely to make the 'best' choice easy (unless the levels of that attribute were all of moderate size and similar on the latent scale). Therefore, the random utility variance, ceteris paribus, would be small compared with a situation with all attributes at intermediate levels. Since the OMEP was a small fraction of the full factorial, it was possible to avoid such problematic situations; the coding of domain levels was chosen to avoid designing situations defined by every attribute at its 'top' level, those with every attribute at its 'bottom' level, and 'easy to choose' situations of the type described above.

Survey structure

The face-to-face interview began with rating questions about respondents' SCRQOL at the time of the survey followed by the discrete choice and best-worst experiments. At the end of the interview, participants provided background information and self-rated their level of

understanding of the experiments. Interviewers also rated respondents' understanding of the experiments.

All respondents participated in both DCE and BWS experiments. In order to control for systematic ordering effects, the order of appearance of DCE and BWS followed the patterns shown in Table 3. In the first pattern, for example, DCE1 appeared first, followed by DCE2 and BWS. Each respondent was assigned one of the ordering sequences at random.
[Table 3, about here]

A key feature of this research involved asking respondents to put themselves in the position of someone who could no longer take care of themselves so that they answer the questions in context. In particular, respondents were asked to imagine that they have had an unspecified accident that has rendered them dependent on others for care. The experiments clearly had the potential to distress some respondents, particularly the more vulnerable. Steps were therefore taken to ensure that respondents were left in a good frame of mind and were not left distressed by taking part in the research. Interviewers were fully briefed on this matter and were given an information leaflet to leave with respondents, providing contact names and telephone numbers they could call if they wished for reassurance.

Data collection

The data collection involved house-to-house recruitment with the questionnaire administered through computer aided personal interview (CAPI). The pilot was conducted in March 2009 among 300 adults in the south east of England and Birmingham. The majority of the interviews were completed within approximately 30 minutes. However, some interviews, mostly among older respondents, took between 45 minutes and 1 hour.

The sample was not specified to be nationally representative, but rather the aim was to over-sample ethnic minorities and those over 65 years of age to allow a more thorough review of

the ability of these population segments to undertake the various choice tasks. This was in line with the objective of using this phase of the research to compare the DCE and BWS approaches to assess how they performed when applied to social care, and how accessible they might be when used with service user groups. Ethical approval for the research and data collection was received from the University of Kent Research Ethics Committee.

Econometric analysis

We used mixed logit models to analyse the DCE and BWS data in order to account for the correlation between observations from the same respondent (i.e., eight in the DCEs and 12 in the BWS). Both the DCE and BWS models captured the main effects of each domain level. Interaction terms were excluded since earlier likelihood-ratio tests and the small correlation among estimated parameters in the final models showed that they were not statistically significant. The following equations present generalised specifications of the DCE1, DCE2 and BWS models, respectively.

- Discrete choice experiments

$$U_{DCE1_{ijt}} = \beta_1 \cdot \text{Food}_{ijt} + \beta_2 \cdot \text{Pers.Care}_{ijt} + \beta_3 \cdot \text{Safety}_{ijt} + \beta_4 \cdot \text{Control}_{ijt} + \beta_5 \cdot \text{Dignity}_{ijt} + \zeta_i + \varepsilon_{ijt}$$

Where $U_{DCE1,ijt}$ represents the utility of an individual *i* of choosing alternative *j* at choice exercise *t*, and $\beta_{1\text{-}5}$ the parameters to be estimated. All variables on the right hand side of the equation (Food, Control, etc.) were dummy coded. For example, in the case of the domain Food, there are three parameters β_{1_2}, β_{1_3}, β_{1_4} corresponding to estimates of the last three levels of the Food domain after setting the parameter of the first level (β_{1_1}) equal to zero;

ζ_i is the error component used to capture the correlation between observations from the same respondent, with mean equal to zero and standard deviation σ_ζ, which is estimated in the model. Finally, ε_{ijt} is the error component due to differences between observations.

Using similar notation, the utility of an individual *i* choosing domain-level *j* at choice exercise t in the DCE2 is given as:

$$\mathrm{U}_{\mathrm{DCE2}_{ijt}} = \beta_6 \cdot \mathrm{Accommodation}_{ijt} + \beta_7 \cdot \mathrm{Soc.Participation}_{ijt} + \beta_8 \cdot \mathrm{Employment}_{ijt} + \beta_9 \cdot \mathrm{Control}_{ijt}$$
$$+\beta_{10} \cdot \mathrm{Dignity}_{ijt} + \beta_{11} \cdot \mathrm{Own\_home}_{ijt} + \lambda_i + \varepsilon_{ijt}$$

where λ is the error component used to capture the correlation between observations from the same respondent (*i*) with mean zero.

- Best-worst scaling

While the choice task in BWS was different; respondent *i* chooses the best and worst aspects of a given situation - this choice task can still be modelled as a discrete choice using a mixed logit model and the operationalisation of BWS data in a random utility model parallels that of a traditional DCE (Flynn, 2010c). Specifically, utility functions were specified for every possible 'best-worst' pair, measuring the difference in the utility between a domain level chosen as the best and another chosen as being worst. Therefore, respondents could choose among 72 'best-worst' pairs in a given situation. For example, if the *Food and Drink* being at level 1 was chosen as best and *Control of over daily life* being at level 4 was chosen as worst within a given situation t, the utility of choosing this 'best-worst' pair would be:

$$\mathrm{U}_{(\text{Food at level 1, Control at level 4})_{ijt}} = [\beta_{\mathrm{Food1}} \cdot (1, \text{if food domain is at level 1; 0 otherwise})$$
$$-[\beta_{\mathrm{Control4}} \cdot (1, \text{if control domain is at level 4, 0 otherwise})] + \omega_i + \varepsilon_{ijt}$$

In common with all limited dependent variable models, the estimates of the attribute level parameters are *not* means: they are a perfect confound of means and variances on the latent utility scale. The variance (more usually conceptualised as its inverse, the variance scale factor) can be 'netted out' by transformation through a numeraire attribute *if and only if* there is no attribute-specific variance heterogeneity (Louviere, 2006), a strong assumption. Whereas in modelling the discrete choice experiments it is necessary to fix one of the levels

of each variable from the equations to avoid over-specification of models, in the case of best-worst this is only necessary for only one level of just one attribute (Flynn et al., 2007).

Results and discussion

Sample

Details of the pilot sample of the 300 respondents who completed the interview are shown in Table 4. Elderly (65+) and ethnic minorities were sufficiently represented in the sample. Out of the 300 participants, 29.3% indicated that they receive benefits or tax credits.

[Table 4, about here]

DCE1 and DCE2 were unlabelled experiments, therefore the prior expectation was to observe a balance in preferences between Situations A and B. In DCE1, Situation A was selected 1,306 times (48%) and Situation B was selected 1,394 times (52%), whereas, in DCE2, Situation A was chosen 1,139 times (47.5%) and Situation B was chosen 1,260 times (52.5%). Preferences between Situation A and B were not statistically different as suggested by the chi-square test for equality of proportions (χ^2=0.405, p = 0.524 > 0.05, *d.f.* = 1). The observed patterns implied that there were no unobserved biases towards a systematic preference for the left or right alternative.

Self-rating questions and interviewers' observations were used to develop criteria for excluding data from further analysis. In particular, respondents should have been able to:

- Put themselves in an imaginary position,
- Understand the descriptions in the choices,
- Look at all aspects of choices,
- Feel that they are able to answer the choices.

In addition, interviewers should have indicated that the respondent:

- Understood the tasks “a little”, “a great deal” or “completely”,
- Gave the questions “some”, “careful” or “very careful” consideration,

- Did not lose concentration in the later stages of the interview.

Finally, the data set in modelling excluded all non-traders; that is respondents who consistently chose situation A or B in all eight exercises in DCE1 or DCE2. Table 5 shows the number of observations available for model development after excluding responses that failed to meet the abovementioned criteria.

[Table 5, about here]

Testing for ordering effects

Prior to the development of the final models, we tested whether the order of appearance of the DCE1, DCE2 and BWS experiments in the survey resulted in significant differences in the models. Estimation results showed that the order of appearance did not affect the level of noise within the responses in all three experiments. More details on these results are available upon request from the first author.

Model estimations

Tables 6, 7 and 8 present the estimated parameters of mixed-logit models using data from the DCE1, DCE2 and the BWS experiments, respectively. In both the DCE1 and DCE2 models, the majority of coefficients were statistically significant at the a=0.05 level and their signs were in line with prior expectations. The trend in the value of the coefficients also follows prior expectations; as the domain level increases (i.e., the situation of a domain becomes worse) the value of the parameter decreases. However, the majority of parameters for the second levels of the domains were not significantly different from zero, implying a non-significant difference in respondents' preference between levels one and two on those domains. This finding suggests that there may be benefit in reviewing some of the wording of the levels for the main study to make them more distinct. Moreover, the coefficients of

dignity[2] and *safety* domains at levels 3 and 4 were not significantly different (i.e., Level 3 = -0.663 ± 1.96*0.12; Level 4 = -0.625 ± 1.96*0.12).

[Table 6, about here]

[Table 7, about here]

The BWS data was used to estimate a single model, which represents the average valuations across all respondents in the sample. Similarly to the results in DCE1 and DCE2, the parameters of level one and two across all domains except *dignity* and the parameters of levels three and four in the *safety* and *dignity* domains were are not statistically different (see Table 8).

[Table 8, about here]

Comparison of values between DCE and BWS

The results from DCE1, DCE2 and the BWS experiments can be compared if we look at the marginal values of moving from the lowest level (e.g. *Dignity_2* in Figure 3a) to the highest level of need (e.g. *Dignity_4* in Figure 3a). In this case, we examine only the differences in preference within a given attribute from the BWS.

It is noteworthy that the models have different scales hence the coefficients cannot be directly compared (Swait & Louviere, 1993), but we can look at the relative size of the differences by using one of the domain levels as a common denominator and scaling all others relative to this. In this case, we have chosen the highest level of need of the control domain (i.e., *Control_4*) which was strongly estimated in all models[3]. Figure 3 provides a comparison of the relative values coming from the different approaches along with standard errors (see Hess & Daly, 2009, for computation of standard errors). In particular, Figure 3a

[2] *Dignity:* Level 3 = -0.663 ±1.96*0.12; Level 4 = -0.625 ± 1.96*0.12; *Safety* Level 3 = -0.662±1.96*0.12; Level 4 = 0.809 ± 1.96*0.11

[3] All negative domain level coefficients have been divided by the negative coefficient of control and that is why the coefficients of each domain level in Figures 3a - 3c appear to be positive.

presents the comparison of domain coefficients that were common across DCE1, DCE2 and BWS. Figure 3b shows the comparison of domain coefficients between BWS and DCE1 and finally, Figure 3c shows the coefficients that appeared in DCE2 against those estimated in BWS.

[Figure 3, about here]

Table 9 presents the results of the statistical comparison across the relative values and standard errors coming from the different approaches. In particular, we used a t-test to examine whether the differences in the rescaled preference weights across DCE1, DCE2, and BWS were statistically different from each other.

[Table 9, about here]

The two approaches did reveal a broadly similar pattern in preferences. In eight of out of 11 cases, the domain-level weights common between DCE1 and BWS were not statistically different at the a=0.05 level. Only, the third level of *Control* and second and third level of *Food and Drink* were statistically different at the 95% and 99.9% confidence intervals, respectively. With regard to domain-level weights that were common in DCE2 and BWS, eight out of 12 weights were not statistically different from each other at the a = 0.05 level of significance. The rescaled BWS estimates are statistically different to those from the DCE2 for *Employment and Occupation* (level 3), *Social Participation* (level 3) and *Living in Own Home*.

Finally, the values placed on *Dignity* and *Control* between the two separate DCE exercises (for which these were the common attributes) were not statistical different. As a result the two different groupings in the design appear to support consistent valuations of these common attributes across respondents.

Conclusions and discussion

This study is the first to compare estimates from DCEs with those from a profile case Best-worst scaling study. Flynn (2010c) discussed the possibility that preference estimates from the two types of task might disagree. However, in this context – that of levels of social care domains that would subsequently be used within a decision making and policy design framework – there are similar patterns in preferences and in the majority of cases the estimated preference weights (when normalised/rescaled) are not statistically different.

The differences that are observed in the BWS and DCE estimates may reflect bias in the latter caused by different respondents inferring different information about the attributes they did not "see": omitting attributes from a DCE typically affects estimates of both the attribute level means and the variance (scale) (Islam, Louviere, & Burke, 2007). The tests for preference equality performed here suggested such effects at the aggregate level may be small. However, there may be important differences among subgroups. Any such differences would have implications for the generalisability of the results to the wider population. Thus, future work on comparing these methods may consider using more sophisticated methods that model heterogeneity, such as the mixed logit, or scale-adjusted latent class analyses (Flynn, 2010b). Alternatively, models that jointly model preference and scale heterogeneity could be estimated, such as generalised multinomial logistic models (G-MNL), which nest both of those models as special cases (Fiebig, Keane, Louviere, & Wasi, 2010).

This study illustrates key issues that are important in choosing between profile case BWS and DCEs studies. The statistical issues in constructing the pairs for the DCE were not pertinent for the profile case BWS study. However, this had to be balanced against the practical issues in coding the BWS design so as to avoid artificially 'easy' or 'difficult' profiles. Presenting single profiles in the BWS study allowed all attributes to be presented. As noted above, this did not appear to lead to differences in patterns of preferences, but the loss of

explicit trade-offs inherent in intra (rather than inter) profile choices mean that the broad equivalence found here may not hold true in other contexts. In the case of the ongoing research on social care outcomes we will take forward the BWS approach as it has a lower cognitive burden and will be more amenable to the collection of preference data from service users.

However, we would also suggest that future studies should conduct piloting to ascertain whether interactions between attributes exist. DCE designs to estimate these interactions are available but in some cases the researcher may be unable to estimate them if using a profile case BWS task (Flynn, 2010b).

It is encouraging that this first comparison of the two tasks has found broad equivalence in estimates. It offers a framework that may be useful to health services researchers needing guidance in choice of discrete choice task in future preference elicitation studies.

Acknowledgements

This project was funded by the NIHR Health Technology Assessment Programme, Project No. 06/96/01. The views and opinions expressed therein are those of the authors alone. Grateful thanks to all those who participated in the research both in the consultation and fieldwork stages; to Accent, which undertook the fieldwork with the general population; and to James Caiels and Diane Fox who undertook many of the cognitive interviews. Last but not least, we would like to thank three anonymous referees for their useful comments

References

Bamford, C., Qureshi, H., Nicholas, E., & Vernon, A. (1999). Outcomes of Social Care for Disabled People and Carers, Outcomes in Community Care Practice, No. 6. York: Social Policy Research Unit, University of York.

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics,* 11, 447-456.

Bliemer, M. C., & Rose, J. (2005). Efficiency and sample size requirements for stated choice studies. (p. 26). Institute of Transport and Logistic Studies, The University of Sydney, Sydney: Institute of Transport and Logistic Studies, The University of Sydney.

Brazier, J., Ratcliffe, J., Salomon, J., & Tsuchiya, A. (2007). Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press.

Burge, P., Netten, A., & Gallo, F. (2010). Estimating the value of social care. *Journal of Health Economics,* 29(6), 883-894.

Caiels, J., Forder, J., Malley, J., Netten, A. and Windle, K. (2010). Measuring the outcome of low-level services: Final Report. Canterbury: PSSRU Discussion Paper, Personal Social Services Research Unit, University of Kent.

Cook, R. D., & Natchtsheim, C. J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics,* 22(315-324).

Fedorov, V. V. (1972). Theory of Optimal Experiments. Translated and edited by W.J. Studden and E.M. Klimko, New York, Academic Press.

Fiebig, D. G., Keane, M. P., Louviere, J., & Wasi, N. (2010). The generalised multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science,* 29(3), 393-421.

Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing,* 11(1), 12-25.

Flynn, T. N. (2010a). Using Conjoint Analysis and Choice Experiments to Estimate QALY Values: Issues to Consider. *Pharmacoeconomics,* 28(9), 711-722.

Flynn, T. N. (2010b). Using discrete choice experiments to understand preferences for quality of life. Variance scale heterogeneity matters. *Social Science and Medicine,* 70, 1957-1965.

Flynn, T. N. (2010c). Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. *Expert Review of Pharmacoeconomics and Outcomes Research,* 10(3), 259-267.

Flynn, T. N., Louviere, J. J., Peters, T. J., & 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.

Glendinning, C., Challis, D., Fernández, J.-L., Jacobs, S., Jones, K., Manthorpe, J., et al. (2008). Evaluation of the Individual Budgets Pilot Programme, Final Report, Social Policy Research Unit, University of York. York.

Grewal, I., Lewis, J., Flynn, T. N., Brown, J., Bond, J., & Coast, J. (2006). Developing attributes for a generic quality of life measure for older people: preference or capabilities. *Social Science and Medicine,* 62, 1891-1901.

Hess, S., & Daly, A. (2009). Calculating errors for measures derived from choice modelling estimates. 88th Annual Meeting of the Transportation Research Board. Washington, D.C. .

Islam, T., Louviere, J. J., & Burke, P. F. (2007). Modelling the effects of including/excluding attributes in choice experiments on systematic and random components. *International Journal of Research in Marketing,* 24, 289-300.

Kuhfeld, W. (2009). Marketing Research Methods in SAS: Experimental Design, Choice, Conjoint and Graphical Techniques. Cary, NC, USA: SAS Institute Inc.

Louviere, J. J. (2006). What you don't know might hurt you: Some unresolved issues in the design and analysis of discrete choice experiments. *Environmental and Resource Economics,* 34(1), 173-188.

Louviere, J. J., & Woodworth, G. G. (1990). Best-Worst Scaling: A Model for Largest Difference Judgments. University of Alberta: Working Paper, Faculty of Business.

Malley, J., Sandhu, S., & Netten, A. (2006). Younger Adults' Understanding of Questions for a Service User Experience Survey, Report to The Health and Social Care Information Centre. Canterbury: PSSRU Discussion Paper 2360, Personal Social Services Research Unit, University of Kent.

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

Miller, E., Cooper, S.-A., Cook, A., & Petch, A. (2008). Outcomes Important to people with intellectual disabilities. *Journal of Policy and Practice in Intellectual Disabilities,* 5(3), 150-158.

Netten, A., Beadle-Brown, J., Trukeschitz, B., Towers, A., Welch, E., Forder, J., et al. (2010). Measuring the Outcomes of Care Homes. Canterbury: PSSRU Discussion Paper 2696/3, Personal Social Services Research Unit, University of Kent.

Netten, A., Malley, J., Forder, J., Burge, P., Potoglou, D., Brazier, J., et al. (2009). Outcomes of Social Care for Adults (OSCA): Interim Report. Canterbury: PSSRU Discussion Paper 2642, Personal Social Services Research Unit, University of Kent.

Netten, A., McDaid, D., Fernández, J.-L., Forder, J., Knapp, M., Matosevic, T., et al. (2005). Measuring and understanding social services outputs. Canterbury: PSSRU Discussion Paper 2132/3, Personal Social Services Research Unit., University of Kent.

Netten, A., Ryan, M., Smith, P., Skatun, D., Healey, A., Knapp, M., et al. (2002). The development of a measure of social care outcome for older people. PSSRU Discussion Paper 1690/2: Personal Social Services Research Unit.

ONS. (2010). Measuring Outcomes for Public Service Users (MOPSU) project: Final report London: Office for National Statistics.

Puckett, S. M., & Rose, J. M. (2009). Observed efficiency of a D-optimal design in an interactive agency choice experiment. In S. Hess & A. Daly (Eds.), International Choice Modelling Conference. Harrogate, UK.

Qureshi, H., Patmore, C., Nicholas, E., & Bamford, C. (1998). Outcomes in Community Care Practice. Overview: Outcomes of Social Care for Older People and Carers. York: Social Policy Research Unit, University of York.

Swait, J., & Louviere, J. J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research,* 30(3), 305-314.

Figure Captions

Fig. 1 Discrete choice experiment situations in the OSCA pilot study

Fig. 2 Best-worst scaling situation from OSCA pilot study

Fig. 3 Comparison of rescaled domain weights (a) BWS and domains common in DCE1 and DCE2; (b) BWS against domains in DCE1, and (c) BWS and domains in DCE2

Which of these two situations would you consider to be the worst?

	Situation A	Situation B
Food and Nutrition	I can't always get all the food and drink I need, and I think there is a risk to my health	I can't always get all the food and drink I need, but I don't think there is a risk to my health
Personal Care	I do not feel adequately clean or presentable	I feel adequately clean and presentable
Safety	Generally I feel as safe as I want	Most of the time I don't feel safe enough
Control over daily life	I have as much control over my daily life as I want	I have as much control over my daily life as I want
The way I am helped	The way I'm helped undermines the way I think and feel about myself	The way I'm helped sometimes undermines the way I think and feel about myself
Choice (mark "X" in worst option)	☐	☐

(a)

Which of these two situations would you consider to be the worst?

	Situation A	Situation B
Accommodation cleanliness and comfort	My home is adequately clean and comfortable	My home is as clean and comfortable as I want
Social participation & involvement	I don't feel lonely and I have enough contact with people I like	I have as much contact as I want with people I like
Use of my time	I spend my time as I want, doing things I value or enjoy	I have enough things I value or enjoy to do with my time
Control over daily life	I have some control over my daily life but not enough	I have some control over my daily life but not enough
The way I am helped	The way I'm helped does not make me think or feel any differently about myself	The way I'm helped does not make me think or feel any differently about myself
Living in own home	And I am living in my own home	And I am not living in my own home
Choice (mark "X" in worst option)	☐	☐

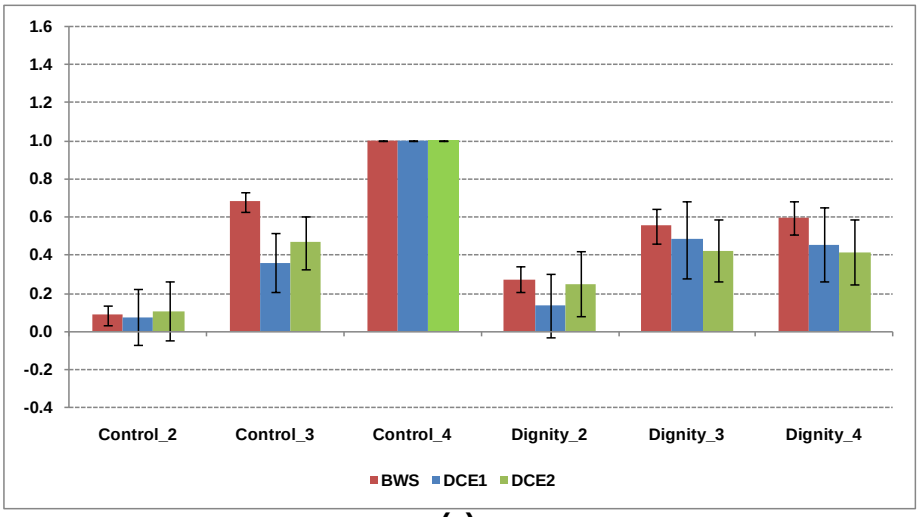
(b)

Fig. 1 Choice situations from OSCA pilot study in (a) DCE 1 and (b) DCE2

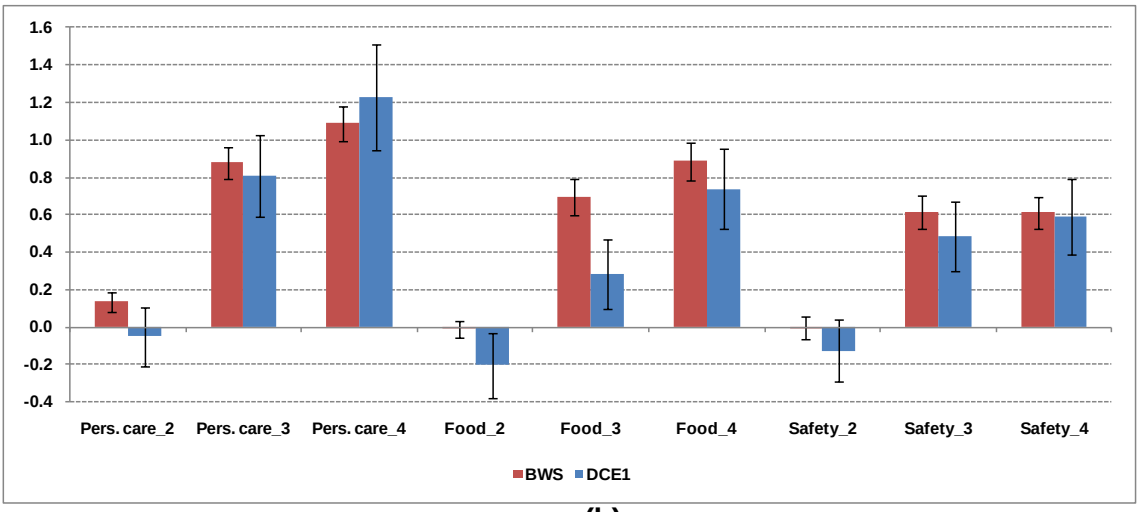
Which of these nine points would rate as being the best and which as being the worst?

Best (mark "X")	Aspect of life	Worst (mark "X")
	I can get all the food and drink I need	
	I have poor personal hygiene, so I don't feel at all clean or presentable	
	My home is as clean and comfortable as I want	
	Sometimes I don't feel safe enough	
	Sometimes I feel lonely, but have some contact with people I like	
	I spend my time as I want, doing things I value or enjoy	
	I have adequate control over my daily life	
	The way I'm helped undermines the way I think and feel about myself	
	And I am living in my own home	

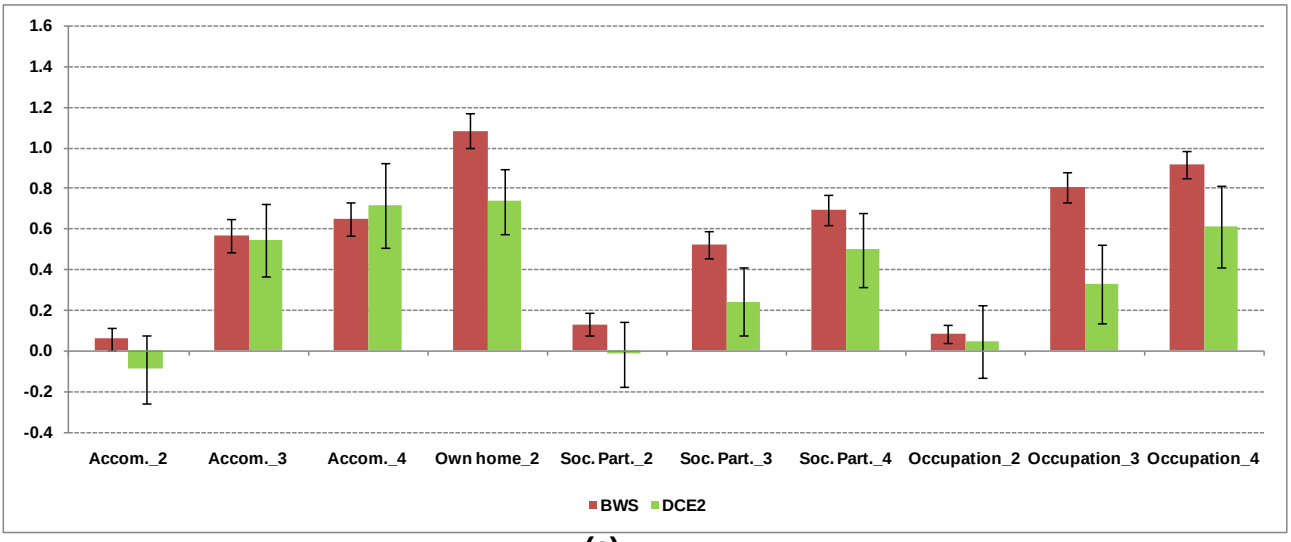
Fig. 2 Best-worst scaling choice exercise from OSCA pilot study



(a)



(b)



(c)

Fig. 3 Comparison of rescaled common domain weights across: (a) BWS, DCE1 and DCE2; (b) BWS and DCE1, and (c) BWS and DCE2[4]

[4] The differences in sign for the ratio of marginal utility of Pers. Care_2, Soc. Part_2 and Accom_2 are not important because the estimated parameters are not statistically different from zero.

Table 1 Social care related quality of life (SCRQOL) domains and levels in the OSCA study

Aspects of SCRQOL	Definition	Level			
		1	2	3	4
Food & Drink	The service user feels he/she has a nutritious, varied and culturally appropriate diet with enough food and drink he/she enjoys at regular and timely intervals	I can get all the food and drink I like when I want	I can get all the food and drink I need	I can't always get all the food and drink I need, but I don't think there is a risk to my health	I can't always get all the food and drink I need, and I think there is a risk to my health
Personal cleanliness and comfort	The service user feels he/she is personally clean and comfortable and looks presentable or, at best, is dressed and groomed in a way that reflects his/her personal preferences	I feel clean and am able to present myself the way I like	I feel adequately clean and presentable	I do not feel adequately clean or presentable	I have poor personal hygiene, so I don't feel at all clean or presentable
Safety	The service user feels safe and secure. This means being free from fear of abuse, falling or other physical harm and fear of being attacked or robbed	Generally I feel as safe as I want	Generally I feel safe enough	Sometimes I don't feel safe enough	Most of the time I don't feel safe enough
Control over daily life	The service user can choose what to do and when to do it, having control over his/her daily life and activities	I have as much control over my daily life as I want	I have adequate control over my daily life	I have some control over my daily life but not enough	I have no control over my daily life
Accommodation cleanliness & comfort	The service user feels their home environment, including all the rooms, is clean and comfortable	My home is as clean and comfortable as I want	My home is adequately clean and comfortable	My home is not very clean or comfortable	My home is not at all clean or comfortable
Social participation & involvement	The service user is content with their social situation, where social situation is taken to mean the sustenance of meaningful relationships with friends, family and feeling involved or part of a community should this be important to the service user	I have as much contact as I want with people I like	I don't feel lonely and I have enough contact with people I like	Sometimes I feel lonely, but have some contact with people I like	Most of the time I feel lonely and very rarely have contact with people I like
Occupation & employment	The service user is sufficiently occupied in a range of meaningful activities whether it be formal employment, unpaid work, caring for others or leisure activities	I spend my time as I want, doing things I value or enjoy	I have enough things I value or enjoy to do with my time	I don't have enough things I value or enjoy with my time	I don't do anything I value or enjoy with my time
Dignity	The psychological impact of support and care on the service user's personal sense of significance	The way I'm helped makes me think and feel better about myself	The way I'm helped does not make me think or feel any differently about myself	The way I'm helped sometimes undermines the way I think and feel about myself	The way I'm helped undermines the way I think and feel about myself
Living in own home	Whether the service user would live in their own home or not	And I am living in my own home	And I am not living in my own home		

Table 2 Grouping of domains between the two choice experiments

DCE1	DCE2
1 Food and drink	**4** Accommodation, cleanliness and comfort
2 Personal care	**5** Social participation and involvement
3 Safety	**6** Employment and occupation (Use of my time)
7 Control over daily life	**7** Control over daily life
8 Dignity (The way I am helped)	**8** Dignity (The way I am helped)
	9 Living in own home

Table 3 Order Patterns of appearance of the experiments

	Order of Appearance		
Pattern	**1st**	**2nd**	**3rd**
1	DCE1	DCE2	BWS
2	DCE2	DCE1	BWS
3	BWS	DCE1	DCE2
4	BWS	DCE2	DCE1

Table 4 Pilot sample

Sample	300
Age 65+ (%)	48.3
Female (%)	48.7
* Age < 65	52.9
* Age ≥ 65	44.1
Ethnic non-white	26.3
Married	56.7
Working Full-Time (%)	22.7
* Age < 65	40.0
* Age ≥ 65	4.1
Retired (%)	50.3
* Age < 65	9.0
* Age ≥ 65	82.9

Table 5 Number of respondents excluded from the discrete choice analysis

Question	DCE1	DCE2	Best-worst
Could not put themselves into an imaginary position	42	42	42
Could not understand the descriptions in the choices	2	2	2
Did not look at all aspects of choices	1	1	1
Felt that they unable to answer the choices	2	2	2
Did not understand very much or at all	3	3	3
Gave the questions little or no consideration	5	5	5
Lost concentration in the later stages	2	2	2
Non-traders	3	5	0
Total number of observations excluded[5]	**60**	**62**	**57**
Total number of observations available for modelling	**240**	**238**	**243**

[5] Numbers for each criterion do not add up to compute the total number of observations excluded as more than one condition may apply to each observation

Table 6 Estimated coefficients in DCE1

Attribute level	Coefficient Value	t-ratio
Dignity		
1. The way I'm helped makes me think and feel better about myself	Reference	
2. The way I'm helped does not make me think or feel any differently about myself	-0.183	-1.6
3. The way I'm helped sometimes undermines the way I think and feel about myself	-0.663	-5.6
4. The way I'm helped undermines the way I think and feel about myself	-0.625	-5.1
Control over daily life		
1. I have as much control over my daily life as I want	Reference	
2. I have adequate control over my daily life	-0.104	-1.0
3. I have some control over my daily life but not enough	-0.495	-4.1
4. I have no control over my daily life	-1.37	-10.5
Safety		
1. Generally I feel as safe as I want	Reference	
2. Generally I feel safe enough	0.172	1.5
3. Sometimes I don't feel safe enough	-0.662	-5.1
4. Most of the time I don't feel safe enough	-0.809	-6.9
Personal care		
1. I feel clean and am able to present myself the way I like	Reference	
2. I feel adequately clean and presentable	0.070	0.6
3. I do not feel adequately clean or presentable	-1.110	-8.8
4. I have poor personal hygiene, so I don't feel at all clean or presentable	-1.680	-10.8
Food and drink		
1. I can get all the food and drink I like when I want	Reference	
2. I can get all the food and drink I need	0.278	2.3
3. I can't always get all the food and drink I need, but I don't think there is a risk to my health	-0.390	-3.0
4. I can't always get all the food and drink I need, and I think there is a risk to my health	-1.010	-8.0
Scale to account for repeated observations	0.112	0.7
Model diagnostics		
Number of observations	240 x 8 = 1920	
D.O.F.	16	
Final log likelihood	-932.4	
$Rho^2(0)$	0.299	

Table 7 Estimated coefficients in DCE2

Attribute level	Coefficient value	t-ratio
Accommodation cleanliness and comfort		
1. My home is as clean and comfortable as I want	Reference	
2. My home is adequately clean and comfortable	0.114	1.0
3. My home is not very clean or comfortable	-0.714	-5.5
4. My home is not at all clean or comfortable	-0.935	-6.8
Social participation and involvement		
1. I have as much contact as I want with people I like	Reference	
2. I don't feel lonely and I have enough contact with people I like	0.019	0.2
3. Sometimes I feel lonely, but have some contact with people I like	-0.318	-3.1
4. Most of the time I feel lonely and very rarely have contact with people I like	-0.651	-4.9
Employment and occupation		
1. I spend my time as I want, doing things I value or enjoy	Reference	
2. I have enough things I value or enjoy to do with my time	-0.061	-0.5
3. I don't have enough things I value or enjoy with my time	-0.430	-4.1
4. I don't do anything I value or enjoy with my time	-0.799	-6.8
Control over daily life		
1. I have as much control over my daily life as I want	Reference	
2. I have adequate control over my daily life	-0.142	-1.2
3. I have some control over my daily life but not enough	-0.607	-5.0
4. I have no control over my daily life	-1.300	-9.0
Dignity		
1. The way I'm helped makes me think and feel better about myself	Reference	
2. The way I'm helped does not make me think or feel any differently about myself	-0.327	-3.2
3. The way I'm helped sometimes undermines the way I think and feel about myself	-0.552	-5.4
4. The way I'm helped undermines the way I think and feel about myself	-0.542	-4.8
Living in own home		
1. And I am living in my own home	Reference	
2. And I am not living in my own home	-0.960	-8.8
Scale to account for repeated observations	0.025	0.18
Model diagnostics		
Number of observations	238 x 8 =1904	
D.O.F.	16	
Final log likelihood	-1031.5	
$Rho^2(0)$	0.218	

Table 8 Estimated coefficients in best-worst scaling

Attribute level	Coefficient value	t-ratio
Dignity		
1. The way I'm helped makes me think and feel better about myself	-1.007	-6.8
2. The way I'm helped does not make me think or feel any differently about myself	-2.327	-14.4
3. The way I'm helped sometimes undermines the way I think and feel about myself	-3.662	-20.1
4. The way I'm helped undermines the way I think and feel about myself	-3.86	-22.0
Control over daily life		
1. I have as much control over my daily life as I want	Reference	
2. I have adequate control over my daily life	-0.421	-3.2
3. I have some control over my daily life but not enough	-3.261	-19.0
4. I have no control over my daily life	-4.798	-24.7
Employment and occupation		
1. I spend my time as I want, doing things I value or enjoy	-0.019	-0.2
2. I have enough things I value or enjoy to do with my time	-0.434	-3.4
3. I don't have enough things I value or enjoy with my time	-3.887	-20.4
4. I don't do anything I value or enjoy with my time	-4.429	-24.0
Social participation and involvement	-0.622	-4.9
1. I have as much contact as I want with people I like	-1.262	-7.9
2. I don't feel lonely and I have enough contact with people I like	-3.139	-16.7
3. Sometimes I feel lonely, but have some contact with people I like	-3.951	-21.2
4. Most of the time I feel lonely and very rarely have contact with people I like	-0.622	-4.9
Safety		
1. Generally I feel as safe as I want	-0.952	-6.2
2. Generally I feel safe enough	-0.937	-5.7
3. Sometimes I don't feel safe enough	-3.895	-20.6
4. Most of the time I don't feel safe enough	-3.886	-21.9
Accommodation cleanliness and comfort		
1. My home is as clean and comfortable as I want	-0.764	-5.2
2. My home is adequately clean and comfortable	-1.059	-6.8
3. My home is not very clean or comfortable	-3.486	-18.6
4. My home is not at all clean or comfortable	-3.880	-20.6
Food and drink		
1. I can get all the food and drink I like when I want	-0.497	-3.3
2. I can get all the food and drink I need	-0.446	-3.0
3. I can't always get all the food and drink I need, but I don't think there is a risk to my health	-3.838	-21.5
4. I can't always get all the food and drink I need, and I think there is a risk to my health	-4.754	-24.5
Personal care		
1. I feel clean and am able to present myself the way I like	-0.165	-1.2
2. I feel adequately clean and presentable	-0.816	-5.4
3. I do not feel adequately clean or presentable	-4.380	-23.2
4. I have poor personal hygiene, so I don't feel at all clean or presentable	-5.384	-27.6
Living in own home		
1. And I am living in my own home	0.322	2.5
2. And I am not living in my own home	-4.883	-27.9
Scale to account for repeated observations	-0.868	-3.4
Model diagnostics		
Number of observations	243 x 12 = 2916	
D.O.F.	34	
Final log likelihood	-8564.5	
$Rho^2(0)$	0.313	

Table 9 Equality test between DCE and BWS preference weights

Comparison	Attribute level	Coefficient (DCE - BWS)	Std. error (DCE- BWS)	t-test	p-value
DCE1 vs. BWS	Pers. care_2	-0.187	0.085	-2.213	p = 0.05
	Pers. care_3	-0.068	0.119	-0.575	n.s.
	Pers. care_4	0.139	0.151	0.917	n.s.
	Food_2	-0.192	0.092	-2.089	p = 0.05
	Food_3	-0.412	0.108	-3.821	p = 0.001
	Food_4	-0.150	0.122	-1.235	n.s.
	Safety_2	-0.122	0.088	-1.383	n.s.
	Safety_3	-0.130	0.105	-1.237	n.s.
	Safety_4	-0.021	0.110	-0.190	n.s.
	Control_2	-0.011	0.080	-0.139	n.s.
	Control_3	-0.319	0.084	-3.795	p = 0.001
	Control_4	0.000			n.s.
	Dignity_2	-0.141	0.091	-1.544	n.s.
	Dignity_3	-0.069	0.113	-0.615	n.s.
	Dignity_4	-0.140	0.109	-1.281	n.s.
DCE2 vs. BWS	Accom._2	-0.149	0.091	-1.650	n.s.
	Accom._3	-0.019	0.100	-0.190	n.s.
	Accom._4	0.069	0.113	0.608	n.s.
	Own home_2	-0.348	0.093	-3.743	p = 0.001
	Soc. Part._2	-0.148	0.086	-1.720	p = 0.1
	Soc. Part._3	-0.281	0.093	-3.006	p = 0.01
	Soc. Part._4	-0.194	0.102	-1.910	p = 0.1
	Occupation_2	-0.040	0.095	-0.418	n.s.
	Occupation_3	-0.476	0.105	-4.515	p = 0.001
	Occupation_4	-0.305	0.109	-2.812	p = 0.01
	Control_2	0.021	0.085	0.247	n.s.
	Control_3	-0.213	0.076	-2.820	p = 0.01
	Control_4	0.000			n.s.
	Dignity_2	-0.024	0.093	-0.255	n.s.
	Dignity_3	-0.129	0.094	-1.366	n.s.
	Dignity_4	-0.180	0.098	-1.835	p = 0.1