Is best-worst scaling suitable for health state valuation? A comparison with discrete choice experiments

KRUCIEN Nicolas
Health Economics Research Unit, University of Aberdeen, Institute of Applied Health Sciences, Foresterhill, Aberdeen, AB25 2QN. United Kingdom, Email: nicolas.krucien@abdn.ac.uk

WATSON Verity
Health Economics Research Unit, University of Aberdeen, Institute of Applied Health Sciences, Foresterhill, Aberdeen, AB25 2QN. United Kingdom, Email: v.watson@abdn.ac.uk

RYAN Mandy
Health Economics Research Unit, University of Aberdeen, Institute of Applied Health Sciences, Foresterhill, Aberdeen, AB25 2QN. United Kingdom, Email: m.ryan@abdn.ac.uk

Corresponding Author
Nicolas KRUCIEN
Health Economics Research Unit
University of Aberdeen
Institute of Applied Health Sciences
Aberdeen, AB25 2QN, UK
Tel: 01224-437892
Fax 01224-437195
Email: nicolas.krucien@abdn.ac.uk

Abstract
Health utility indices (HUIs) are widely used in economic evaluation. The best-worst scaling (BWS) method is being used to value dimensions of HUIs. However, little is known about the properties of this method. This paper investigates the validity of the BWS method to develop HUI, comparing it to another ordinal valuation method, the discrete choice experiment (DCE). Using a parametric approach we find a low level of concordance between the two methods, with evidence of preference reversals. BWS responses are subject to decision biases, with significant effects on individuals' preferences. Non parametric tests indicate BWS data has lower stability, monotonicity and continuity compared to DCE data, suggesting the BWS provides lower quality data. As a consequence, for both theoretical and technical reasons, practitioners should be cautious both about using the BWS method to measure health-related preferences, and using HUI based on BWS data. Given existing evidence it seems that the DCE method is a better method, at least because its limitations (and measurement properties) have been extensively researched.

Keywords
Best-worst scaling; Discrete choice experiment; Health utility; Stated preferences

JEL codes
C35; C99; D01; I10

1. Introduction

Economic evaluation informs resource allocation in health care systems around the world. For instance, the National Institute for Health & Care Excellence (NICE) typically recommend interventions for reimbursement if a one unit improvement in Quality Adjusted Life Years costs less than £30,000 (NICE 2013). However, economic evaluation raises questions about the measurement and valuation of health care benefits: *What should be considered as a "health benefit"? How should different levels of benefit be valued?* There is consensus that health-related quality of life (HRQoL) is a multi-dimensional construct. Many health utility indices (HUIs) have been developed to value the multi-dimensional nature of HRQoL (e.g. EQ-5D, SF-6D, ICECAP-O, ICECAP-A).

To develop HUIs (health) dimensions and (quality) levels are combined to form health states or profiles and utility values are generated for these profiles. The validity of cardinal techniques for valuing profiles, such as time trade-off (TTO) and standard gamble (SG) has been debated (Tijhuis 2000; Green, Brazier, and Deverill 2000; Arnold et al. 2009), leading to the use of ordinal methods such as discrete choice experiments (DCEs) and best-worst scaling (BWS) (Ryan, Gerard, and Amaya-Amaya 2008; T. N. Flynn et al. 2007). These ordinal techniques seem well-suited to investigate the multi-dimensional nature of HRQoL because they are themselves based on multi-attribute choice mechanisms (Lancaster 1966).

DCEs were introduced into health economics in the early 1990s to value benefits beyond health outcomes (de Bekker-Grob, Ryan, and Gerard 2012; Ryan and Gerard 2003), and present individuals with a number of choices sets, each of which contain two or more profiles that vary with respect to attribute levels. For each choice, respondents are assumed to make trade-offs between attributes; based on these trade-offs they state what profile they would choose. From individuals' responses information is obtained on the utility of the attributes and how individuals make trade-offs across attributes (marginal rates of substitution, MRS). Many studies have investigated the extent to which DCE responses satisfy the underlying theoretical axioms and provided researchers with knowledge about the method's limitations and what can be done to overcome them (San Miguel, Ryan, and Scott 2002; Ryan and San Miguel 2003; Miguel, Ryan, and Amaya-Amaya 2005; Bryan et al. 2000; Ryan and Bate 2001; McIntosh and Ryan 2002; Bryan and Dolan 2004; Scott 2002). Although DCEs were introduced to value attributes beyond health outcomes, there is growing interest in using the method to develop HUIs. DCEs have been applied to develop programme specific HUIs in social care for the elderly(Ryan et al. 2006), glaucoma related health states (Burr et al. 2007), social care (Potoglou et al. 2011), and generic HUIs including the EQ-5D (Bansback et al. 2012; Viney et al. 2014; Norman, Cronin, and Viney 2013).

Since its introduction to health economics in 2007, one variant of the BWS method, known as BWS case 2 or profile-case, has been used to develop HUIs, particularly HUIs implementing a capabilities approach to valuation (e.g., ICECAP-A, ICECAP-O, CHU-9D)

(Al-Janabi, Flynn, and Coast 2010; T. N. Flynn et al. 2013; Coast et al. 2008; Ratcliffe et al. 2012; Potoglou et al. 2011) (See supplementary material 1 for a summary of the application of BWS case 2 in health). The stated appeal of BWS is its simplicity and ability to measure "more preferences" than the DCE approach. A BWS survey includes multiple choice tasks; each choice presents individuals with a health state (profile) and asks them to select the "best" and "worst" features of the profile. The repetition of these best and worst decisions over multiple tasks allows weights to be estimated for each feature, and thus a (utility) score for defined health profiles to be calculated.

Importantly, in both DCE and BWS methods the coefficients/utility weights estimated in HUIs are assumed to represent individuals' trade-offs between dimensions. It is this assumption that allows utility to be estimated for different profiles. DCE tasks ask respondents to make such trade-offs. However, it is not clear that the BWS method elicits or can be used to infer trade-offs. Further, as with DCEs, other choice-based methods such as TTO and SG also ask participants to choose between competing health states and then to perform between-profile comparisons. However, the BWS method only asks individuals to make within-profile choices; and opportunity cost is not embedded within this task. This is a fundamental difference between BWS and other valuation techniques; in the welfarist tradition only choices between alternatives, and implied trade-offs, provide the necessary information to derive utility-based preferences (Mooney 2009). As noted by Coast et al (2008), in a BWS experiment only "*values and not preferences are elicited*", and the authors distinguish between values and preferences "*because individuals are not asked to trade one thing for another*" (Coast et al. 2008). There is thus a mismatch between the information elicited in BWS tasks and their use to develop HUIs. Furthermore, BWS does not provide information about the absolute value of dimensions; a participant might select a feature as being best, but this does not necessary mean that she/he positively values the feature. These two limitations mean that information obtained from BWS choices might not be suitable for measurement of HRQoL preferences.

In this study we compare the stated preferences (HRQoL weights) and measurement properties of BWS case 2 and DCE methods when used to develop a HUI. The stated preference comparison indicates whether the two methods lead to same results and therefore whether the methods are substitutes. If the BWS and DCE methods lead to different results, it is important to determine which method is the most appropriate for HUI development. To do this the measurement properties (i.e., stability, monotonicity, continuity, completeness) of the methods are compared. **We focus on the BWS case 2 (hereafter refer to as BWS) for three reasons: (1) It** is the most widely applied BWS case in health economics; (2) It has been used mainly for developing HUIs; (3) It has been presented as an alternative to DCE method (T. N. Flynn et al. 2007). In comparison, BWS case 1 is an alternative to rating scale methods and BWS case 3 is an extended DCE (Lancsar et al. 2013; J. Louviere et al. 2013; Lee, Soutar, and Louviere 2008).

Three studies have compared DCE and BWS results and have mixed findings. Whilst Potoglou et al (2011) and Flynn, Peters and Coast (2013) find broadly similar results, Whitty et al (2014) find low agreement between the two methods (Potoglou et al. 2011; Flynn, Peters, and Coast 2013; Whitty et al. 2014). Whitty et al (2014) find that (1) 72% of participants prefer the DCE method, (2) DCE choices are more consistent (in a repeated choice task 23% of DCE respondents and 10% of BWS respondents made the same choices), and (3) DCE respondents are more likely to adopt a compensatory choice behaviour (i.e., to make trade-offs among attributes). According to van Dijk et al (2013), the BWS tasks were perceived as being more difficult and took longer to complete than DCE tasks (van Dijk et al. 2013). This mixed evidence suggests that the "validity and acceptability of the BWS method is not definite and requires further research.

Our study contributes to this literature in two ways. First it provides new empirical evidence on the comparison between DCEs and BWS within the context of developing a HUI. Although, BWS has been predominately used in health economics to develop HUIs, previous comparison studies have elicited preferences in other contexts (healthcare priority setting; social care; preferences for capabilities based quality of life instrument). We compare the methods when developing a HUI for glaucoma. Second, we test the measurement properties of BWS. Whilst much evidence exists about the measurement properties of DCEs, such research has not been conducted for BWS.

2. Experimental design

2.1. Context

The DCE and BWS methods were used to develop a Glaucoma Utility Index. Detailed information on the study design are available in Burr et al (2007)(Burr et al. 2007). Six attributes described glaucoma-related health states: *central and near vision* (Vision); *lighting and glare* (Light); *mobility* (Mobility); *activities of daily living* (Daily); *local eye discomfort* (Eye); *other effects of Glaucoma and its treatment* (Other). Each attribute was described by the same four levels: *No difficulty (1); Some difficulty (2); Quite a lot of difficulty (3); Severe difficulty (*4).

2.2. The questionnaire

Experimental design methods (i.e. orthogonal main effects plan and its foldover) resulted in 32 DCE choice tasks. Each task was a choice between two generic health states (A vs B). Participants were asked to select the worst health state (Figure 1). The BWS tasks were generated using the same orthogonal main effects plan, resulting in 32 BWS tasks (corresponding to alternative A of the DCE tasks). For each BWS task, respondents were asked to select the 'best' and 'worst' attribute levels. The first (or left) alternative in the DCE task was always the BWS task. This means that the alternative presented in the BWS task did not always correspond to the least preferred alternative in the DCE task.

Each participant was asked to answer the DCE task first, immediately followed by the corresponding BWS task (DCE #1 was followed by BWS #1; DCE #2 followed by BWS #2; (…); DCE #36 followed by BWS #36). This approach reduced the impact of learning & fatigue effects on the comparison of DCE and BWS performance[1]. Three versions of the questionnaire, with 8, 16 or 32 "pages" (each page included both the DCE and BWS task) were tested with results indicating that respondents were able to handle 32 pages (based on response rates, item response rates, and rationality tests)(Burr et al. 2007). In addition to the 32 experimental tasks, two tasks were repeated to test the stability of choices (task 6 was repeated as task 28 and task 12 as the task 13) and two warm-up tasks were also included, thus resulting in 36 DCE and BWS tasks (See Supplementary material 2 for a full list of choice tasks). The order in which the attributes were presented was sorted by increasing level of severity, thus varying the order attributes' across choice tasks for both BWS and DCEs[2].

2.3. Subject recruitment and ethics

The study used a within-subject design in which 293 patients completed both the DCE and BWS tasks in a self-administered paper-based questionnaire. It was not possible to perform a formal sample size computation as we had no prior information about participants' preferences for the different attribute levels. An approximate formula from Louviere et al (2000) indicates that a minimum of 73 and 178 participants were needed for the DCE and BWS methods respectively (See appendix 1 for sample size computation). Respondents were selected from patients at four hospital-based and one community-based glaucoma clinics across the United Kingdom. Respondents were also recruited from The International Glaucoma Association (IGA). Ethical approval for the study was obtained from the Central Office of Research Ethics Committees. The research was conducted according to the tenets of the Declaration of Helsinki.

3: Do the BWS and DCE methods lead to same preferences for HUIs?

In this section we compare stated preferences between the DCE and BWS methods after accounting for a number of behavioural effects likely to influence how participants answer the choice questions. Both the DCE and BWS data are typically modelled within the random utility framework using a multinomial logit (MNL) model (See appendix 2 for detailed explanation of the econometric framework), which assumes: (1) respondents share similar tastes (no variability in tastes), (2) have the same ability to choose (no variability in choice

[1] In the DCE literature choice consistency (as approximated by errors variance) is found to follow a U-shaped relationship with error variance decreasing at the beginning of the sequence of tasks, presumably as a result of "learning" effect, and then stabilising before increasing again towards the end of the sequence of tasks, presumably as a result of a "fatigue" effect. Moreover learning and fatigue effects don't' appear to be symmetrical (i.e., larger learning effect). If participants first answered all the DCE (or BWS) tasks, and then all the BWS (or DCE) ones, we would confound true differences in responses to the different tasks with the learning/fatigue effects due to the task order.

[2] As noted by one of the reviewers, altering the order of the attributes' levels across the choice tasks but within the participants might have some detrimental effects on the quality of the data. Readers interested in this specific issue could refer to Appendix 5 of this study which provides a detailed discussion and analysis of the issue.

consistency) and (3) make *bias-free* decisions (decisions not influenced by ordering effects). We relax these behavioural assumptions by successively estimating three choice models:

- The 1st model is a multinomial logit model (MNL) allowing for decision biases in participants' choices. Decision biases refer to any systematic effects of factors other than attributes' levels that impact on choice (e.g. location of attribute levels within the choice options). In our model we investigate two potential decision biases: (1) Order biases (effect of attribute location within the alternative); and (2) Attribute biases (systematic preference for one attribute regardless of its value). Previous studies found evidence of ordering effects (e.g., left option is more likely to be selected) in DCE tasks (Kjaer et al. 2006). Similarly, one could expect top-located attributes to be more likely to be selected (either as best or worst) in the BWS tasks.
- The second model is a random parameters logit or mixed multinomial logit (MMNL) model, which is typically used to explore inter-individual variability in preferences by describing individuals' preferences with probability distributions (McFadden and Train 2000).
- The third model is a generalised multinomial logit (GMNL) model. In any choice model, estimated preferences ($\hat{\beta}$) perfectly confound participants' *true* preferences (β^*) with error variance (σ_ε). The GMNL model has been used to quantify the variability in respondents' choices that is *likely to be due* to underlying changes in (σ_ε) rather than (β^*) (Fiebig et al. 2010a; Keane and Wasi 2012). In the choice modelling literature, (σ_ε) is sometimes used as a proxy measure for "ability to choose", that is error variance increases as participants' choice behaviour becomes more random (or less consistent).

We estimate these models to account for behavioural differences in DCE and BWS responses that may confound the comparison of stated preferences. First, it has been argued that BWS tasks are easier to answer than the DCE tasks (T. N. Flynn et al. 2007). Therefore BWS choices would be less influenced by inter-individual differences in response to task difficulty (as approximated by variability in choice ability (σ_ε)). Second, if respondents differ in the way they interpret "best" and "worst" then answers to the BWS tasks will be more diverse than answers to DCE tasks. These inter-individual differences in understanding of the BWS questions could lead to more choice variability in BWS tasks than DCE tasks. Third, respondents are less familiar with the tasks of rank ordering of product features than choosing among competing products. Therefore, we expect answers to BWS questions to be more influenced by decision biases, such as order effects.

3.1: Parametric analysis of DCE method

Given the DCE included six dimensions, each with four levels, 18 preference parameters ($\beta_{2:19}$) were estimated. For each dimension, the *a priori* most desirable level (no difficulty) was used as the reference level for dummy coding. We specified an order bias (ORD) parameter (β_1) to capture the systematic effect of the first (left) alternative.

$$V_{njt} = \beta_1 ORD_{jt} + \beta_{2:4} Daily[2:4]_{jt} + \beta_{5:7} Vision[2:4]_{jt} + \beta_{8:10} Other[2:4]_{jt} + \beta_{11:13} Light[2:4]_{jt} + \beta_{14:16} Eye[2:4]_{jt} + \beta_{17:19} Mobility[2:4]_{jt}$$

3.2: Parametric analysis of BWS method

For the BWS method, only one attribute level is used as the reference point allowing estimation of up to 23 preference parameters ($\beta_{11:33}$). The level "severe difficulty" for the dimension "other effects of glaucoma and its treatment" was selected as the reference point. This level was identified as the least attractive feature using a count analysis of the best and worst choices (See supplementary material 3 for results of the count analysis). We included two other sets of variables to investigate decision biases. We used five order variables (ORD[2:6]) to capture the systematic effect on individuals' decisions of an attribute's location within the health profiles ($\beta_{1:5}$). We included five attribute variables (ATT) to capture a systematic effect of the HUI attributes on individuals' decisions ($\beta_{6:10}$), regardless of the (quality) levels used to describe the attributes. Such decision bias would be in line with a simplifying heuristic according to which respondents would only pay attention to the HUI dimensions rather than their levels.

$$V_{njt} = \beta_{1:5} ORD[2:6]_{jt} + \beta_6 ATT_{Daily\,jt} + \beta_7 ATT_{Vision\,jt} + \beta_8 ATT_{Other\,jt} + \beta_9 ATT_{Light\,jt} + \beta_{10} ATT_{Eye\,jt} + \beta_{11:14} Daily[1:4]_{jt} + \beta_{15:18} Vision[1:4]_{jt} + \beta_{19:21} Other[1:3]_{jt} + \beta_{22:25} Light[1:4]_{jt} + \beta_{26:29} Eye[1:4]_{jt} + \beta_{30:33} Mobility[1:4]_{jt}$$

3.3: Comparison of preferences between BWS and DCE

The estimated preferences ($\hat{\beta}$) are not directly comparable across the BWS and DCE methods because of differences in their measurement scales (Louviere et al. 2002). First, the two methods measure preferences on scales with different origin (reference) points. The DCE measures preferences for $\sum_k (L_k - 1)$ attributes' levels, where (L_k) indicates the number of levels for attribute (k), relative to (K) reference levels. For example, the preferences for the "Vision" attribute are estimated for three levels *Vision_Some, Vision_Quite, Vision_Severe* relative to *Vision_No*. In the BWS method, preferences are relative to one common origin point $[(\sum_k L_k) - 1]$. Preferences for the "Vision" attributes are estimated for four levels *Vision_No, Vision_Some, Vision_Quite, Vision_Severe* relative to *Other_Severe*. This difference in the origin points requires rescaling to ensure estimates are comparable. Second, estimated preferences ($\hat{\beta}$) perfectly confound *true* preferences (β^*) with the errors variance (σ_ε^2) therefore differences between the DCE and BWS results may be due to differences in *true* preferences, model scale or a combination of both. The error variance confound is usually neutralised by computing MRS. We use the same approach and divide the rescaled estimates by a common denominator (See appendix 3 for detailed rescaling procedure of the DCE and BWS estimates).

The rescaled estimates ($\hat{b}$) for the DCE and BWS methods are then compared using both Pearson (r) and Kendall (τ) correlation coefficients. We plot the rescaled estimates and fit a linear trend (i.e. $\hat{b}_{BWS} = \mu + \alpha . \hat{b}_{DCE}$) to investigate: (i) the coefficient of determination (R^2) indicating the overall agreement (matching) between the two methods; (ii) the slope

parameter (α) which can be interpreted as a scale effect (e.g., α=2 would indicate that BWS estimates are consistently double those of the DCE) and (iii) the intercept parameter (μ) which quantifies the bias in the estimates of one method relative to the other (e.g., μ=0.5 would indicate that BWS estimates are the same as DCE estimates after 0.5 has been added to them).

3.4: Results of the parametric analyses

Table 1 presents the results of the choice models. Across the DCE models most preference parameters (except "*Other: Some*" and "*Other: Quite*") are statistically significant at the 5% confidence level with the expected sign. Across the BWS models, only one preference parameter ("*Light: No*") is not significant. The best fitting model for both the DCE and BWS methods is the GMNL model.

In all three models we find evidence of decision biases in both the DCE and BWS data. Regarding the DCE, the left-to-right order bias is negative and statistically significant, indicating that respondents were less likely to select the left alternative (alternative A) all other things being equal. Regarding the BWS data, the presentation order of the attributes in the BWS task [ORD] is statistically significant and consistent with a central fixation bias i.e. the two middle attributes (i.e., ORD.3, ORD.4) are more likely to be selected both as best and worst than the top (i.e., ORD.1, ORD.2) and bottom (i.e., ORD.5, ORD.6) attributes. We find that four attributes (i.e., Daily, Vision, Other and Light) were more likely to be selected both as best and worst relative to the "Mobility" attribute. The amount of bias varies between attributes with statistically significant differences in some cases, and the implied ordering of the attributes and therefore the ranking of health state values differs across model specifications (See supplementary material 4 for comparison of BWS results with and without decision biases). Together the results on order and attribute effects indicate that BWS responses are prone to decision biases that should be accounted for when estimating preferences for HRQoL dimensions.

The comparisons of rescaled estimates across the BWS and DCE methods for the three choice models indicate *weak* to *moderate* correlations between the BWS and DCE results. The Spearman correlations are 0.593 ($P_{0.05}$ = 0.014), 0.610 ($P_{0.05}$ = 0.011) and 0.272 ($P_{0.05}$ = 0.029) for the MNL allowing for decision biases, MMNL and GMNL models respectively. The Figure 2 presents a visual comparison of the rescaled estimates. The results indicate that the linear trend is only weakly supported (cf. R^2 values), suggesting a low level of matching between the BWS and DCE methods. Regarding the GMNL model, there is no difference of scaling between the two methods (α=1.034), but there is a substantial bias (γ=0.490).

Our results also indicate that the BWS method would generate attribute level estimates that are closer to each other. For example, regarding the "Mobility" attribute, with the DCE method the rescaled estimate for the highest level of severity is 3 times bigger than rescaled estimate for 2nd lowest level of severity (= 0.900/0.293). A similar comparison for the BWS method leads to a 1.5 times difference (=0.45/0.31). For all the remaining attributes, except

"Light", we find a similar pattern (DCE vs BWS): "Eye" (1.25 vs 0.48); "Light" (0.68 vs 1.10); "Other" (5.55 vs 2.16); "Vision" (3.63 vs 0.83); "Daily" (3.38 vs 2.51).

We find evidence of large differences in the results of the DCE and BWS methods. The differences between the two methods could be attributed to differences in accuracy (i.e. one method would systematically misestimate participants' preferences) rather than precision. This has important implications for the development of HUIs because the choice of one method or the other will lead to a different weighting of the HRQoL dimensions with evidence of preference reversals.

4: Which of the methods is the best to develop HUIs?
In this section, we compare the two methods on four conditions derived from micro-economic consumer theory: stability; monotonicity; continuity; and completeness.

4.1: Stability
Stability refers to the ability of respondents to make the same choices when confronted with the same tasks. Here two choice tasks are repeated - task #6 repeated as task #28 and task #12 repeated as task #13. *Strong* and *weak* versions of the stability test are defined - the strong test is based on the distant tasks (i.e., task #6 vs task #28) and the weak version is based on the adjacent tasks (i.e., task #12 vs task #13). For the DCE, stability is satisfied when the respondent makes the same choice in the two tasks; for the BWS stability is satisfied when the respondent makes the same best or worst choice in both tasks.

Table 2 presents the results of the stability tests. For both methods the proportions of respondents failing the stability test is not significantly different across the 'weak' and 'strong' version of the stability test (McNemar: DCE_$P_{0.05}$ = 0.256; BWS_$P_{0.05}$ = 0.105). 13% (14/106) and 24% (58/243) of the respondents fail both stability tests for the DCE and BWS methods respectively. Across the two methods we find significant differences in the proportion of respondents failing the stability tests (McNemar: Strong_$P_{0.05}$ = 0.017; Weak_$P_{0.05}$ = 0.024). Answers to the BWS questions are less stable than DCE ones.

4.2: Monotonicity
Monotonicity implies more of a desirable feature and less of an undesirable feature is preferred. We test monotonicity using dominance tasks in which participants are expected to make a particular decision if they hold monotonic preferences for the dimensions. For the DCE method, the dominance tasks are choice tasks in which one alternative is more attractive (or dominates) than the other, and respondents fail monotonicity when they select the dominated alternative. We apply this test to five DCE tasks {7; 18; 21; 27; 32}. For example in task #7, the sequence of attribute levels for alternatives A and B are respectively {113311} and {224422}, thus alternative A outperforms B on every attribute.

To the best of our knowledge, monotonicity of BWS responses has not previously been tested and we propose a testing procedure (See appendix 4 for details). We modify the above test for application in BWS. We identify six tasks {6; 17; 19; 28; 34; 36} where either the best or worst choice is obvious. For example in BWS task #19, the attributes' levels are {443144}, and then we expect respondents to select the 4th attribute as best since it is associated with no difficulty compared to all other attributes offering quite a lot or severe difficulty

Table 3 presents the results. Regarding the DCE method, 73% of respondents *fully satisfied* monotonicity (i.e., for all 5 tasks) and only 2% *fully failed* the monotonicity test. In comparison the BWS method leads to poorer results, with 0% of respondents *fully satisfying* the monotonicity test and 42% of respondents who *fully failed* the test.

4.3: Continuity

Continuity refers to the assumption that respondents are making trade-offs across attributes, implying compensatory decision making in which a deterioration in one attribute can be compensated for by an improvement in another. Continuity has been investigated in the DCE literature by looking at the proportion of choices based on one attribute (Scott 2002; Ryan and Bate 2001; McIntosh and Ryan 2002; Ryan, Watson, and Entwistle 2009). When the respondent considers only one attribute (i.e. case of perfect lexicographic or dominant preference), their lexicographic score is 100% (i.e. in 100% of the tasks the respondent always selects the alternative with the highest level of the attribute). In contrast, when the respondent makes *random* choices, the lexicographic score is expected to be approximatively 50% (i.e., in 50% of the tasks the respondent selects the alternative with the highest level of the attribute). Respondents are assumed to exhibit dominant preference for a particular attribute when the lexicographic score is larger than 90%. We test the continuity requirement is done at the individual level by computing a lexicographic score for each respondent. A similar approach is used for the BWS method. The lexicographic score varies between 0% and 100%. A score of 100% would indicate that the respondent always selects a given attribute, either as Best or Worst, in all tasks. Alternatively, a score of 0% would indicate that the respondent never selects the attribute, either as Best or Worst, in any of the tasks. The lexicographic score would be approximately 17% in case of *random* choice behaviour. Respondents are assumed to exhibit dominant preference for a given attribute when the lexicographic score is larger than 50%.

In the DCE, 16.6% of respondents (43/259) have a dominant preference for one attribute. For 66% of these respondents (28/43), this lexicographic preference is for the "Vision" attribute. In the BWS, 23.5% of respondents (61/259) have a dominant preference for one attribute, and in 67% of the cases (41/61) have a dominant preference for the "Vision" attribute too. These proportions of respondents with non-continuous preferences are significantly different between the DCE and BWS methods (McNemar: $P_{0.05} = 0.047$), with the BWS performing worse than DCE.

4.4: Completeness

Completeness is defined as the ability of respondents to make a choice according to a rank ordering of available options. We test completeness indirectly by testing the ability of the DCE and BWS methods to generate a (strict) ordering of attribute levels (Lagerkvist 2013). This ordering is based on the probability (P_{ij}) of an attribute level *i* being preferred to another attribute level *j*:

$$P_{ij} = \left(\frac{A + B}{A + 2B + C}\right) 100$$

where A corresponds to the number of times attribute level *i* is ranked better than attribute level *j*; B is the number of ties in ranks, and (C) the number of times attribute level *j* is ranked better than attribute level *i*. For the DCE, preferences are estimated at the individual level, and the estimates for each respondent are used to obtain a rank-ordering of the attribute levels. The rank-orders are used to compute A, B, and C, and the attribute levels are ranked by the sum of (P_{ij}) probabilities. Two attribute levels are considered as belonging to the same rank when there is no significant difference between their probabilities. For the BWS, the observed best and worst decisions are used to directly derive a rank order of the attribute levels for each respondent. Again the individual rankings are used to compute A, B, and C for each pair of attributes' levels. The rankings obtained from the DCE and BWS methods are compared using the Kendall (τ) coefficient of correlation.

Table 4 presents the results of the completeness tests. Both methods use the same number of ranks (i.e. 15 out of the 24 possible in case of perfect discriminatory choices) to order the attributes levels with the same level of dispersion in the rankings, meaning that DCE and BWS methods perform similarly in terms of preference completeness. However the rankings from the two methods are uncorrelated (Kendall correlation: Tau-b=-0.222, $P_{0.05}$ = 0.1406) suggesting thus that the methods provide different insights into patients' preferences with some cases of preference reversals (e.g. no difficulty for Mobility attribute).

4.5 Summary of measurement properties

In summary, we find that BWS method performs worse than the DCE in terms of stability (Failure: 24% vs 13%), monotonicity (Failure: 42% vs 2%) and continuity (Failure: 23% vs 16%). The two methods perform equally well in terms of completeness, but in different ways.

5. Discussion

In this study we addressed two questions "*Do the BWS and DCE methods lead to same preferences for HUIs?*" and "***Which method would provide the best HRQoL measures?"*** To answer the first question we compared the preferences obtained from the two methods using choice models that make different assumptions about participants' choice behaviour. Overall we find a low level of agreement between the results of the DCE and BWS methods. The differences in preferences for HRQoL dimensions are mainly attributable to a change in accuracy rather than precision. Our results are consistent with Whitty et al (2014) who found a similar lack of agreement in personal preferences for a health outcome.

We find that the BWS method tends to generate less differentiated weights, supporting only partially a monotonic function of HUI levels. In Potoglou et al (2011), the BWS method would generate more dispersed weights for 3 attributes (i.e., *Social participation and involvement; Personal care; Employment and occupation*) and less dispersed weights for 3 attributes (i.e., *Safety; Accommodation cleanliness and comfort; Food and drink*). In Whitty et al (2014), weights dispersion can be analysed for two attributes (out of 7) and the BWS method would generate more dispersed weights for one attribute (i.e., *Age*) and less dispersed weights for the other attribute (i.e., *Number of patients*). In our study, less dispersed weights implies that everything else being equal the BWS method would place less value on "large" improvements in different HRQoL dimensions compared to the DCE method. This result has important consequences for the economic evaluation of health technologies as closer utility values for different health states imply a smaller effect of an intervention and a smaller cost-utility measure.

We answer the second research question by comparing the measurement properties of the DCE and BWS. We find that the BWS method performs worse than the DCE method in terms of stability, monotonicity and continuity. This suggests that the HRQoL weights obtained with the BWS method would have lower quality compared those obtained with the DCE method. Whitty et al (2014) also compare the stability of DCE and BWS choices using a repeated task. For the DCE method, they find 75% consistency; similarly, we find 77% for our *weak* stability tests. However, in our study we obtain different results for the BWS choices. Whilst we find that respectively 3% and 66% of the best and worst choices are consistent, Whitty et al (2014) find 64% of best choices and 49% of worst choices are consistent. This discrepancy might be explained by the topic of the studies. Whitty et al (2014) investigated public preferences for healthcare priority setting. In our study, we explore patients' preferences for glaucoma-related quality of life dimensions. All the attributes' levels were negatively framed (i.e. they correspond to different severity/impairment levels), potentially making the worst decisions more relevant.

This study is not exempt from limitations. First, all choice tasks were completed in a fixed order and BWS task appeared immediately after its corresponding DCE task. This approach was used to reduce the impact of learning & fatigue effects on the comparison of DCE and BWS performance. This format may introduce anchoring effects in the comparisons if respondents try to be consistent. However, this would increase similarities across the two methods and reduce our ability to detect differences.

Second, respondents were presented with a large number of tasks (36 DCE + 36 BWS tasks). This may have introduced biases and differences if respondent fatigue and simplifying heuristics affect BWS and DCE decisions differently. However, we expect this effect to have a weak influence on our results. Initially three different versions of the questionnaire were piloted, including 8, 16 and 32 "pages" (BWS + DCE tasks) respectively. The comparisons of

response rates and answers to the stability and dominance tests across the three versions indicated that participants were able to handle 32 pages. Evidence from other DCE studies also suggest that increases in the number of choice tasks has a detrimental effect only for large number of tasks (i.e., > 60) (Bech, Kjaer, and Lauridsen 2011; Louviere et al. 2013).

Third, our stability results are largely influenced by how the test is specified. Due to differences in the composition of the choice sets (i.e., choice among two alternatives for the DCE and choice among six items for the BWS), the chance of successfully passing the stability test by *chance* was 25% for the DCE and 5% for the BWS method. This was not accounted for when comparing the two methods, because the lower chance of providing stable choices was seen as a property of the BWS method. *Ceteris paribus* participants would be less likely to be consistent in their best and worst decisions.

Fourth, our monotonicity results for the BWS method rely on how the analysis is carried out. To the best of our knowledge our study is the first to investigate this measurement property for the BWS method and we develop our own procedure, which needs to be replicated/tested in other studies. Our proposed approach offers a structured framework to check the quality of the BW choices and it can be further refined if additional information about the perceived importance and value of the different attributes & levels is collected by the researcher. As suggested by one of the reviewers, including information from the dominated tasks could lead to biased parameter estimates. However additional analyses (see Appendix 6) indicate that such bias would be negligible.

Finally, we acknowledge we used "ill-defined" choice questions for the BWS method (i.e., *participants were asked to select the best and worst aspects of the health profiles*). This ill-defined format might introduce an "interpretation bias" in the BWS data. For example, some participants might have interpreted best/worst as "most/least worrying features" and others as "most/least desirable features". However, our description of the choice tasks within the BWS was purposeful as the objective of the study was to compare the DCE and BWS methods as they have been used in the health economics literature (Potoglou et al. 2011; T. N. Flynn et al. 2008). Future research could employ latent class approaches to identify different patterns of BW choices that could be linked to potential differences in the understanding of the best and worst concepts.

6. Conclusion

In line with recent evidence, our study results suggest that for both theoretical and technical reasons practitioners should be careful in using the BWS method to measure health preferences. Given existing knowledge the DCE method is likely to be a better method, at least because its limitations have been extensively researched. A comprehensive research

programme examining the strengths and limitations of the BWS methods should be conducted before it is used as an alternative to DCEs.

Acknowledgement

The University of Aberdeen (UoA) and the Chief Scientist Office (CSO) of the Scottish Government Health and Social Care Directorates fund the Health Economics Research Unit (HERU). We thank all participants who took part in the study and the WH Ross foundation that supported the data collection. We also thank authors of the original study (Mary Kilonzo, Jennifer Burr and Luke Vale) for their contribution to questionnaire design and data collection. The views expressed in this paper are those of the authors only and not those of the funding bodies.

References

Al-Janabi, H., Terry N. Flynn, and Joanna Coast. 2010. “Estimation of a Preference-Based Carer Experience Scale.” *Medical Decision Making* 31 (3): 458–68. doi:10.1177/0272989X10381280.

Arnold, D., A. Girling, A. Stevens, and R. Lilford. 2009. “Comparison of Direct and Indirect Methods of Estimating Health State Utilities for Resource Allocation: Review and Empirical Analysis.” *BMJ* 339 (jul20 3): b2688–b2688. doi:10.1136/bmj.b2688.

Bansback, Nick, John Brazier, Aki Tsuchiya, and Aslam Anis. 2012. “Using a Discrete Choice Experiment to Estimate Health State Utility Values.” *Journal of Health Economics* 31 (1): 306–18. doi:10.1016/j.jhealeco.2011.11.004.

Bech, Mickael, Trine Kjaer, and Jørgen Lauridsen. 2011. “Does the Number of Choice Sets Matter? Results from a Web Survey Applying a Discrete Choice Experiment.” *Health Economics* 20 (3): 273–86. doi:10.1002/hec.1587.

Bryan, Stirling, and Paul Dolan. 2004. “Discrete Choice Experiments in Health Economics. For Better or for Worse?” *The European Journal of Health Economics: HEPAC: Health Economics in Prevention and Care* 5 (3): 199–202. doi:10.1007/s10198-004-0241-6.

Bryan, Stirling, Lisa Gold, Rob Sheldon, and Martin Buxton. 2000. “Preference Measurement Using Conjoint Methods: An Empirical Investigation of Reliability.” *Health Economics* 9 (5): 385–95. doi:10.1002/1099-1050(200007)9:5<385::AID-HEC533>3.0.CO;2-W.

Burr, Jennifer M., Mary Kilonzo, Luke Vale, and Mandy Ryan. 2007. “Developing a Preference-Based Glaucoma Utility Index Using a Discrete Choice Experiment.” *Optometry and Vision Science* 84 (8): 797–808.

Caussade, Sebastián, Juan de Dios Ortúzar, Luis I. Rizzi, and David A. Hensher. 2005. “Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates.” *Transportation Research Part B: Methodological* 39 (7): 621–40. doi:10.1016/j.trb.2004.07.006.

Coast, Joanna, Terry N. Flynn, Lucy Natarajan, Kerry Sproston, Jane Lewis, Jordan J. Louviere, and Tim J. Peters. 2008. “Valuing the ICECAP Capability Index for Older People.” *Social Science & Medicine* 67 (5): 874–82. doi:10.1016/j.socscimed.2008.05.015.

de Bekker-Grob, Esther W., Mandy Ryan, and Karen Gerard. 2012. “Discrete Choice Experiments in Health Economics: A Review of the Literature.” *Health Economics* 21 (2): 145–72. doi:10.1002/hec.1697.

DeShazo, J.R., and German Fermo. 2002. “Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency.” *Journal of Environmental Economics and Management* 44 (1): 123–43. doi:10.1006/jeem.2001.1199.

Fiebig, Denzil G., Michael P. Keane, Jordan Louviere, and Nada Wasi. 2010a. “The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity.” *Marketing Science* 29 (3): 393–421. doi:10.1287/mksc.1090.0508.

———. 2010b. “The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity.” *Marketing Science* 29 (3): 393–421. doi:10.1287/mksc.1090.0508.

Flynn, Terry N., Elisabeth Huynh, Tim J. Peters, Hareth Al-Janabi, Sam Clemens, Alison Moody, and Joanna Coast. 2013. “Scoring the ICECAP-A Capability Instrument. Estimation of a UK General Population Tariff.” *Health Economics (Online Preview)*. doi:10.1002/hec.3014.

Flynn, Terry N., Jordan J. Louviere, Tim J. Peters, and Joanna Coast. 2007. “Best-Worst Scaling: What It Can Do for Health Care Research and How to Do It.” *Journal of Health Economics* 26 (1): 171–89. doi:10.1016/j.jhealeco.2006.04.002.

Flynn, Terry N, Jordan J Louviere, Tim J Peters, and Joanna Coast. 2008. “Estimating Preferences for a Dermatology Consultation Using Best-Worst Scaling: Comparison of Various Methods of Analysis.” *BMC Medical Research Methodology* 8 (1): 76. doi:10.1186/1471-2288-8-76.

Flynn, Terry, Tim Peters, and Joanna Coast. 2013. “Quantifying Response Shift or Adaptation Effects in Quality of Life by Synthesising Best-Worst Scaling and Discrete Choice Data.” *Journal of Choice Modelling* 6: 34–43. doi:10.1016/j.jocm.2013.04.004.

Green, Colin, John Brazier, and Mark Deverill. 2000. “Valuing Health-Related Quality of Life. A Review of Health State Valuation Techniques.” *PharmacoEconomics* 17 (2): 151–65.

Hess, Stephane, and John M. Rose. 2012. “Can Scale and Coefficient Heterogeneity Be Separated in Random Coefficients Models?” *Transportation* 39 (6): 1225–39. doi:10.1007/s11116-012-9394-9.

Hsieh, Chang-ming. 2012. “Should We Give Up Domain Importance Weighting in QoL Measures?” *Social Indicators Research* 108 (1): 99–109. doi:10.1007/s11205-011-9868-8.

Keane, Michael, and Nada Wasi. 2012. “COMPARING ALTERNATIVE MODELS OF HETEROGENEITY IN CONSUMER CHOICE BEHAVIOR: MODELS OF HETEROGENEITY IN CONSUMER CHOICE BEHAVIOR.” *Journal of Applied Econometrics*, September, n/a – n/a. doi:10.1002/jae.2304.

Kjaer, Trine, Mickael Bech, Dorte Gyrd-Hansen, and Kristian Hart-Hansen. 2006. “Ordering Effect and Price Sensitivity in Discrete Choice Experiments: Need We Worry?” *Health Economics* 15 (11): 1217–28. doi:10.1002/hec.1117.

Lagerkvist, Carl Johan. 2013. “Consumer Preferences for Food Labelling Attributes: Comparing Direct Ranking and Best–worst Scaling for Measurement of Attribute Importance, Preference Intensity and Attribute Dominance.” *Food Quality and Preference* 29 (2): 77–88. doi:10.1016/j.foodqual.2013.02.005.

Lancaster, Kelvin J. 1966. “A New Approach to Consumer Theory.” *Journal of Political Economy* 74 (2): 132–57.

Lancsar, Emily, Jordan Louviere, Cam Donaldson, Gillian Currie, and Leonie Burgess. 2013. “Best Worst Discrete Choice Experiments in Health: Methods and an Application.” *Social Science & Medicine* 76: 74–82. doi:10.1016/j.socscimed.2012.10.007.

Lee, Julie Anne, Geoffrey Soutar, and Jordan Louviere. 2008. “The Best–Worst Scaling Approach: An Alternative to Schwartz’s Values Survey.” *Journal of Personality Assessment* 90 (4): 335–47. doi:10.1080/00223890802107925.

Louviere, Jordan J., Richard T. Carson, Leonie Burgess, Deborah Street, and A.A.J. Marley. 2013. “Sequential Preference Questions Factors Influencing Completion Rates and Response Times Using an Online Panel.” *Journal of Choice Modelling* 8 (September): 19–31. doi:10.1016/j.jocm.2013.04.009.

Louviere, Jordan, Ian Lings, Towhidul Islam, Siegfried Gudergan, and Terry Flynn. 2013. “An Introduction to the Application of (case 1) Best–worst Scaling in Marketing Research.” *International Journal of Research in Marketing* 30 (3): 292–303. doi:10.1016/j.ijresmar.2012.10.002.

Louviere, Jordan, Deborah Street, Richard Carson, Andrew Ainslie, J.R. DeShazo, Trudy Cameron, David Hensher, Robert Kohn, and Tony Marley. 2002. “Dissecting the Random Component of Utility.” *Marketing Letters* 13 (3): 177–93. doi:10.1023/A:1020258402210.

Mas-Colell, Andreu. 1995. *Microeconomic Theory*. New York: Oxford University Press.

McFadden, Daniel. 1974. “Conditional Logit Analysis of Qualitative Choice Behavior.” In *Frontier in Econometrics*, 105–42. New York: Academic Press.

McFadden, Daniel, and Kenneth Train. 2000. “Mixed MNL Models for Discrete Response.” *Journal of Applied Econometrics* 15 (5): 447–70. doi:10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

McIntosh, E., and M. Ryan. 2002. “Using Discrete Choice Experiments to Derive Welfare Estimates for the Provision of Elective Surgery: Implications of Discontinuous Preferences.” *Journal of Economic Psychology* 23 (3): 367–82. doi:10.1016/S0167-4870(02)00081-8.

Miguel, Fernando San, Mandy Ryan, and Mabelle Amaya-Amaya. 2005. “‘Irrational’ Stated Preferences: A Quantitative and Qualitative Investigation.” *Health Economics* 14 (3): 307–22. doi:10.1002/hec.912.

Mooney, Gavin H. 2009. *Challenging Health Economics*. Oxford ; New York: Oxford University Press.

NICE. 2013. "Guide to the Methods of Technology Appraisal 2013." The National Institute for Health and Care Excellence. http://www.nice.org.uk/article/pmg9/resources/non-guidance-guide-to-the-methods-of-technology-appraisal-2013-pdf.

Norman, Richard, Paula Cronin, and Rosalie Viney. 2013. "A Pilot Discrete Choice Experiment to Explore Preferences for EQ-5D-5L Health States." *Applied Health Economics and Health Policy* 11 (3): 287–98. doi:10.1007/s40258-013-0035-z.

Potoglou, Dimitris, Peter Burge, Terry Flynn, Ann Netten, Juliette Malley, Julien Forder, and John E. Brazier. 2011. "Best–worst Scaling vs. Discrete Choice Experiments: An Empirical Comparison Using Social Care Data." *Social Science & Medicine* 72 (10): 1717–27. doi:10.1016/j.socscimed.2011.03.027.

Ratcliffe, Julie, Terry Flynn, Frances Terlich, Katherine Stevens, John Brazier, and Michael Sawyer. 2012. "Developing Adolescent-Specific Health State Values for Economic Evaluation: An Application of Profile Case Best-Worst Scaling to the Child Health Utility 9D." *PharmacoEconomics* 30 (8): 713–27. doi:10.2165/11597900-000000000-00000.

Ryan, Mandy, and Angela Bate. 2001. "Testing the Assumptions of Rationality, Continuity and Symmetry When Applying Discrete Choice Experiments in Health Care." *Applied Economics Letters* 8 (1): 59–63. doi:10.1080/135048501750041312.

Ryan, Mandy, and Karen Gerard. 2003. "Using Discrete Choice Experiments to Value Health Care Programmes Current Practice and Future Research Reflections." *Applied Health Economics and Health Policy* 2 (1): 55–64.

Ryan, Mandy, Karen Gerard, and Mabel Amaya-Amaya, eds. 2008. *Using Discrete Choice Experiments to Value Health and Health Care*. The Economics of Non-Market Goods and Resources 11. Dordrecht: Springer.

Ryan, Mandy, Ann Netten, Diane Skåtun, and Paul Smith. 2006. "Using Discrete Choice Experiments to Estimate a Preference-Based Measure of outcome—An Application to Social Care for Older People." *Journal of Health Economics* 25 (5): 927–44. doi:10.1016/j.jhealeco.2006.01.001.

Ryan, Mandy, and Fernando San Miguel. 2003. "Revisiting the Axiom of Completeness in Health Care." *Health Economics* 12 (4): 295–307. doi:10.1002/hec.730.

Ryan, Mandy, Verity Watson, and Vikki Entwistle. 2009. "Rationalising the 'irrational': A Think Aloud Study of Discrete Choice Experiment Responses." *Health Economics* 18 (3): 321–36. doi:10.1002/hec.1369.

San Miguel, Fernando, Mandy Ryan, and Anthony Scott. 2002. "Are Preferences Stable? The Case of Health Care." *Journal of Economic Behavior & Organization* 48 (1): 1–14. doi:10.1016/S0167-2681(01)00220-7.

Scott, Anthony. 2002. "Identifying and Analysing Dominant Preferences in Discrete Choice Experiments: An Application in Health Care." *Journal of Economic Psychology* 23 (3): 383–98. doi:10.1016/S0167-4870(02)00082-X.

Tijhuis, G J. 2000. "Value of the Time Trade off Method for Measuring Utilities in Patients with Rheumatoid Arthritis." *Annals of the Rheumatic Diseases* 59 (11): 892–97. doi:10.1136/ard.59.11.892.

Train, Kenneth. 2009. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge ; New York: Cambridge University Press.

Trauer, T., and A. Mackinnon. 2001. "Why Are We Weighting? The Role of Importance Ratings in Quality of Life Measurement." *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 10 (7): 579–85.

van Dijk, J.D., K.G. Groothuis-Oudshoorn, D. Marshall, and M.J. IJzerman. 2013. "Empirical Comparison Of Discrete Choice Experiment And Best-Worst Scaling To Estimate Stakeholders' Risk Tolerance For Hip Replacement Surgery." *Value in Health* 16 (3): A226–27. doi:10.1016/j.jval.2013.03.1149.

Viney, Rosalie, Richard Norman, John Brazier, Paula Cronin, Madeleine T. King, Julie Ratcliffe, and Deborah Street. 2014. “AN AUSTRALIAN DISCRETE CHOICE EXPERIMENT TO VALUE EQ-5D HEALTH STATES: AN AUSTRALIAN DISCRETE CHOICE EXPERIMENT TO VALUE EQ-5D HEALTH STATES.” *Health Economics* 23 (6): 729–42. doi:10.1002/hec.2953.

Whitty, J, J Ratcliffe, G Chen, and PA Scuffham. 2014. “Australian Public Preferences for the Funding of New Health Technologies: A Comparison of Discrete Choice and Profile Case Best-Worst Scaling Methods.” *Medical Decision Making* 34 (5): 638–54. doi:10.1177/0272989X14526640.

FIGURES & TABLES

Figure 1. Illustration of DCE-BWS choice tasks

Choice 1 **Which situation is the worse for you?**	
SITUATION A	SITUATION B
No difficulty with: • Central and near vision • Activities of daily living **Some difficulty** with: • Lighting and glare • The effects of glaucoma and its treatments **Quite a lot of difficulty** with: • Mobility **Severe difficulty** with: • Local eye discomfort	**No difficulty** with: • Local eye discomfort **Some difficulty** with: • Central and near vision • Activities of daily living **Quite a lot of difficulty** with: • Lighting and glare • The effects of glaucoma and its treatments **Severe difficulty** with: • Mobility
(Tick one box only) ☐ **Situation A**	☐ **Situation B**

Remember: tick just **one** aspect as best aspect and **one** as worst aspect

Best aspect	Aspects of situation A	*Worst aspect*
	No difficulty with central and near vision	
	No difficulty with activities of daily living	
	Some difficulty with lighting and glare	
	Some difficulty with the effects of glaucoma and its treatments	
	Quite a lot of difficulty with mobility	
	Severe difficulty with local eye discomfort	

Figure 2. Comparison of rescaled estimates between the BWS and DCE methods

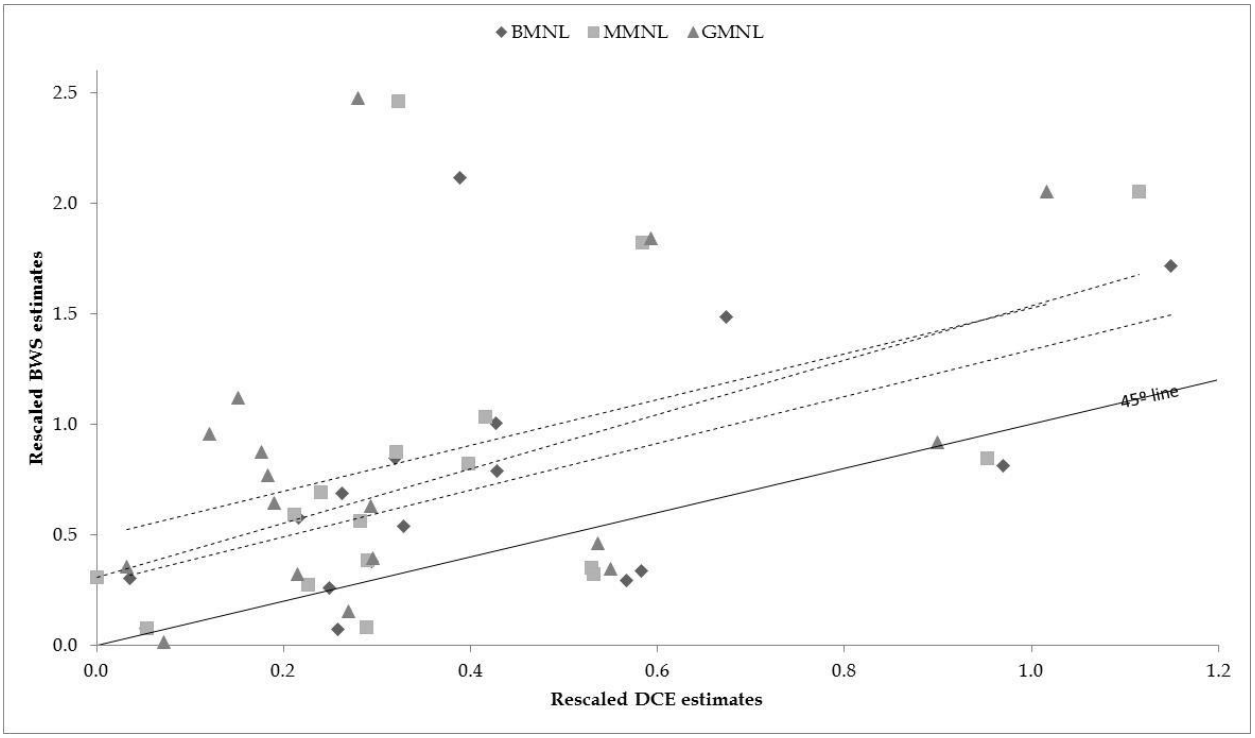


Table 1 - Part 1/2. Estimates (and standard errors) of choice models*

Parameter	MNL		MMNL		GMNL	
	DCE	BWS	DCE	BWS	DCE	BWS
1. Mean preference parameters						
Daily.No	0	**0.717 (0.065)**	0	**0.71 (0.083)**	0	**0.714 (0.067)**
Daily.Some	**0.292 (0.042)**	**0.204 (0.066)**	**0.369 (0.051)**	**0.198 (0.076)**	**0.503 (0.064)**	**0.21 (0.062)**
Daily.Quite	**0.576 (0.051)**	**1.171 (0.064)**	**0.672 (0.062)**	**1.174 (0.071)**	**0.936 (0.09)**	**1.158 (0.067)**
Daily.Severe	**0.99 (0.042)**	**2.05 (0.072)**	**1.27 (0.057)**	**2.048 (0.087)**	**1.702 (0.123)**	**1.979 (0.083)**
Vision.No	0	**3.283 (0.071)**	0	**3.739 (0.098)**	0	**3.638 (0.121)**
Vision.Some	**0.385 (0.045)**	**0.461 (0.065)**	**0.41 (0.055)**	**0.45 (0.075)**	**0.476 (0.066)**	**0.505 (0.064)**
Vision.Quite	**0.667 (0.051)**	**1.298 (0.065)**	**0.742 (0.062)**	**1.304 (0.071)**	**1.009 (0.091)**	**1.304 (0.07)**
Vision.Severe	**1.138 (0.042)**	**0.992 (0.063)**	**1.416 (0.056)**	**0.996 (0.076)**	**1.728 (0.136)**	**1.036 (0.067)**
Other.No	0	**0.921 (0.064)**	0	**0.921 (0.076)**	0	**0.977 (0.068)**
Other.Some	0.035 (0.042)	**0.513 (0.062)**	0.001 (0.051)	**0.511 (0.074)**	-0.056 (0.055)	**0.524 (0.061)**
Other.Quite	0.052 (0.051)	**1.024 (0.069)**	0.068 (0.062)	**1.024 (0.072)**	0.123 (0.068)	**1.002 (0.071)**
Other.Severe	**0.261 (0.043)**	0	**0.305 (0.053)**	0	**0.311 (0.072)**	0
Light.No	0	-0.004 (0.065)	0	-0.039 (0.081)	0	-0.113 (0.063)
Light.Some	**0.424 (0.048)**	**1.054 (0.061)**	**0.504 (0.059)**	**1.058 (0.069)**	**0.301 (0.069)**	**0.997 (0.064)**
Light.Quite	**0.423 (0.051)**	**1.337 (0.063)**	**0.528 (0.064)**	**1.339 (0.072)**	**0.259 (0.077)**	**1.31 (0.068)**
Light.Severe	**0.316 (0.043)**	**1.129 (0.068)**	**0.407 (0.053)**	**1.131 (0.078)**	**0.206 (0.068)**	**1.103 (0.07)**
Eye.No	0	**1.089 (0.064)**	0	**1.083 (0.075)**	0	**1.026 (0.069)**
Eye.Some	**0.246 (0.045)**	**1.439 (0.062)**	**0.288 (0.055)**	**1.444 (0.073)**	**0.365 (0.067)**	**1.439 (0.069)**
Eye.Quite	**0.214 (0.051)**	**1.865 (0.066)**	**0.268 (0.061)**	**1.869 (0.08)**	**0.324 (0.073)**	**1.843 (0.078)**
Eye.Severe	**0.256 (0.045)**	**1.189 (0.072)**	**0.367 (0.054)**	**1.189 (0.084)**	**0.458 (0.083)**	**1.224 (0.076)**
Mobility.No	0	-0.129 (0.067)	0	**-0.162 (0.079)**	0	**-0.243 (0.066)**
Mobility.Some	**0.325 (0.044)**	**0.596 (0.064)**	**0.358 (0.053)**	**0.589 (0.074)**	**0.498 (0.065)**	**0.558 (0.063)**
Mobility.Quite	**0.562 (0.05)**	**0.267 (0.065)**	**0.676 (0.061)**	**0.266 (0.083)**	**0.912 (0.083)**	**0.345 (0.062)**
Mobility.Severe	**0.96 (0.044)**	**0.961 (0.07)**	**1.211 (0.059)**	**0.969 (0.079)**	**1.531 (0.116)**	**0.923 (0.071)**
2. Standard deviation preference parameters						
Daily.No	-	-	0	0.171 (0.143)	0	**0.235 (0.051)**
Daily.Some	-	-	0.026 (0.16)	0.004 (0.374)	0.004 (0.074)	0.034 (0.046)
Daily.Quite	-	-	0.048 (0.198)	0.046 (0.296)	0.016 (0.085)	0.027 (0.07)
Daily.Severe	-	-	**0.836 (0.062)**	0.134 (0.144)	**0.712 (0.078)**	0.049 (0.075)
Vision.No	-	-	0	**1.432 (0.085)**	0	**0.561 (0.096)**
Vision.No	-	-	0.008 (0.183)	0.062 (0.177)	0.027 (0.098)	0.107 (0.056)
Vision.Quite	-	-	0.205 (0.128)	0.019 (0.21)	0.213 (0.118)	0.006 (0.043)
Vision.Severe	-	-	**1.218 (0.06)**	0.094 (0.179)	**1.133 (0.111)**	**0.188 (0.043)**
Other.No	-	-	0	**0.245 (0.104)**	0	**0.279 (0.051)**
Other.Some	-	-	0.073 (0.12)	0.059 (0.242)	0.038 (0.082)	0.001 (0.055)
Other.Quite	-	-	0.068 (0.175)	0.052 (0.256)	0.001 (0.108)	**0.121 (0.053)**
Other.Severe	-	-	**0.597 (0.068)**	0	**0.666 (0.075)**	0
Light.No	-	-	0	**0.476 (0.08)**	0	**0.144 (0.05)**
Light.Some	-	-	0.012 (0.135)	0.012 (0.237)	0.079 (0.114)	**0.104 (0.043)**
Light.Quite	-	-	0.034 (0.13)	0.017 (0.239)	0.126 (0.09)	0.001 (0.042)
Light.Severe	-	-	**0.365 (0.08)**	0.014 (0.256)	**0.44 (0.077)**	0.106 (0.071)
Eye.No	-	-	0	**0.328 (0.098)**	0	**0.277 (0.058)**
Eye.Some	-	-	0.075 (0.132)	0.023 (0.223)	0.027 (0.099)	0.038 (0.042)
Eye.Quite	-	-	0.069 (0.147)	0.013 (0.25)	0.029 (0.09)	0.004 (0.045)
Eye.Severe	-	-	**0.529 (0.069)**	0.138 (0.138)	**0.554 (0.066)**	**0.206 (0.057)**

** In bold, parameters significant at 5% level*

Table 1 - Part 2/2. Estimates (and standard errors) of choice models*

Parameter	MNL		MMNL		GMNL	
	DCE	BWS	DCE	BWS	DCE	BWS
2. Standard deviation preference parameters (continuation)						
Mobility.No	-	-	0	**0.588 (0.071)**	0	**0.219 (0.054)**
Mobility.Some	-	-	0.003 (0.147)	0.014 (0.183)	0.002 (0.068)	0.011 (0.053)
Mobility.Quite	-	-	0.022 (0.086)	0.007 (0.259)	0.016 (0.068)	0.072 (0.066)
Mobility.Severe	-	-	**0.881 (0.062)**	0.183 (0.166)	**0.807 (0.078)**	0.1 (0.103)
3. Heterogeneity parameters (GMNL)						
Tau (τ)	-	-	-	-	**0.839 (0.065)**	**0.437 (0.03)**
Gamma (γ)	-	-	-	-	**0.664 (0.122)**	**4.114 (0.599)**
4. Decision bias parameters						
ASC.A	**-0.447 (0.025)**	-	**-0.528 (0.031)**	-	**-0.54 (0.03)**	-
ATT.Mobility	-	0	-	0	-	0
ATT.Daily	-	**0.191 (0.032)**	-	**0.209 (0.035)**	-	**0.21 (0.033)**
ATT.Vision	-	**0.161 (0.032)**	-	**0.153 (0.035)**	-	**0.144 (0.033)**
ATT.Other	-	**0.285 (0.032)**	-	**0.295 (0.035)**	-	**0.291 (0.033)**
ATT.Light	-	**0.17 (0.03)**	-	**0.181 (0.033)**	-	**0.181 (0.031)**
ATT.Eye	-	0.053 (0.031)	-	0.06 (0.034)	-	0.053 (0.032)
ORD.1	-	0	-	0	-	0
ORD.2	-	**-0.382 (0.031)**	-	**-0.343 (0.036)**	-	**-0.328 (0.032)**
ORD.3	-	**0.177 (0.028)**	-	**0.234 (0.033)**	-	**0.261 (0.029)**
ORD.4	-	**0.349 (0.028)**	-	**0.408 (0.034)**	-	**0.438 (0.029)**
ORD.5	-	**-0.602 (0.034)**	-	**-0.544 (0.037)**	-	**-0.516 (0.035)**
ORD.6	-	**-0.393 (0.033)**	-	**-0.336 (0.041)**	-	**-0.322 (0.035)**
5. Model statistics						
# Observations	9,037	18,074	9,037	18,074	9,037	18,074
# Parameters	19	33	37	56	39	58
Log-likelihood	-5,140.2	-27,805.2	-4,759.3	-27,574.0	-4,678.6	-27,458.3
BIC	10,453.5	55,934.0	9,855.6	55,696.8	9,712.5	55,485.1

** In bold, parameters significant at 5% level*

Table 2. Descriptive analysis of the stability of the BWS and DCE choices

Method	Test	N	Pass	Fail	% Pass	% Fail
BWS *	*Both*	243	139	58	57%	24%
DCE	*Both*	106	54	14	51%	13%
BWS (Best only)	*Both*	247	2	233	1%	94%
BWS (Worst only)	*Both*	245	138	65	56%	27%
BWS *	*Strong*	247	157	90	64%	36%
DCE	*Strong*	124	79	45	64%	36%
BWS (Best only)	*Strong*	250	7	243	3%	97%
BWS (Worst only)	*Strong*	249	154	95	62%	38%
BWS *	*Weak*	258	173	85	67%	33%
DCE	*Weak*	228	176	52	77%	23%
BWS (Best only)	*Weak*	261	9	252	3%	97%
BWS (Worst only)	*Weak*	259	171	88	66%	34%

** For the BWS method, the stability condition is satisfied when at least one choice (either best or worst) is repeated.*

Table 3. Comparative analysis of the monotonicity condition

Nbr of tests successfully passed	DCE (N=245)	BWS (N=235)
0	2%	42%
1	1%	25%
2	4%	32%
3	5%	1%
4	14%	0%
5	73%	0%
6	-	0%

Table 4. Comparison of the completeness results between the methods

Attribute	Level	DCE rank	BWS rank	Absolute difference
Daily	No	1	11	10
Daily	Some	4	14	10
Daily	Quite	12	8	4
Daily	Severe	15	3	12
Vision	No	2	1	1
Vision	Some	2	12	10
Vision	Quite	12	6	6
Vision	Severe	14	8	6
Other	No	8	10	2
Other	Some	6	13	7
Other	Quite	6	10	4
Other	Severe	11	13	2
Light	No	3	14	11
Light	Some	11	9	2
Light	Quite	10	5	5
Light	Severe	9	7	2
Eye	No	6	7	1
Eye	Some	7	4	3
Eye	Quite	8	2	6
Eye	Severe	10	8	2
Mobility	No	1	15	14
Mobility	Some	5	12	7
Mobility	Quite	13	15	2
Mobility	Severe	14	10	4
Measures of dispersion:				
Median Absolute Deviation		5.19	4.45	-
Coefficient of Variation		55.1	45.48	-

APPENDICES

Appendix 1: Sample size computation

Louviere et al (2000) approximate formulae for choice proportions:

$$n \geq \frac{(1-p)}{Tp(a^2)}\left[\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right]^2$$

where (Φ) represents the normal cumulative distribution function. Given a chance probability of 50% (p=0.5) for the DCE method and 16.7% for the BWS method, accuracy level of 90% (a=0.1), confidence level of 95% (α=0.05), 32 and 64 observations per participant (T=32) respectively for the DCE and BWS method, an expected response rate of 20% and 10% exclusion rate (mainly due to missing values), we needed to recruit at least 73 and 178 participants respectively for the DCE and BWS method.

Appendix 2: Econometric framework for modelling of DCE and BWS data

For all four models, the utility (U) that individual ($n=1,\ldots,N$) derives from the health profile ($j=1,\ldots,J$) at the choice occasion ($t=1,\ldots,T$) is a latent variable made up a systematic – observable – component (V) and a stochastic – unobservable – component (ε) (Equation 1). The systematic part of the utility is defined by the combination of individuals' preferences (β_n) and HUI dimensions ($X_{k=1,\ldots,K}$) which is typically assumed to be additive and linear (Equation 2). Given ε, one can only predict the probability (P) of making a specific decision (y) assuming that individuals try to maximise their utility. Assuming ε is independently and identically distributed (IID) as a type 1 extreme value (EV1), the observed choices can be analysed with logit models (McFadden 1974; Train 2009) (Equation 3).

$$U_{njt} = V_{njt} + \varepsilon_{njt} \qquad (1)$$

$$V_{njt} = \sum_k \beta_{nk} X_{tjk} \qquad (2)$$

$$\varepsilon_{njt} \sim iid\ EV1$$

$$P_{njt} = P(y_{nt} = j) = \frac{exp(V_{njt})}{\sum_j exp(V_{njt})} \qquad (3)$$

In the MNL model, all individuals are assumed to hold similar preferences, i.e. $\beta_{nk} = \beta_k\ \forall(n;k)$. Both MMNL and GMNL models relax this constraint, allowing preferences to be distributed over the sample (Equation 4). The individuals' preferences (β_{nk}) are typically assumed to be multi-variate normally (MVN) distributed, with a mean vector (μ_K) and covariance matrix (Ω_K) (Equation 5) where (I_K) refers to the identity matrix of dimension (K) and (L) is the Cholesky factor of (Ω) such that LL'=Ω.

$$\beta_{nk} = \mu_k + L\sigma_{nk} \qquad (4)$$

$$\sigma_{nk} \sim MVN(0, \Omega_k) \qquad (5)$$

$$\Omega_K = \sigma_K I_K \qquad (6)$$

The GMNL model makes this specification more flexible by including two other parameters, (τ_n) and (γ), which control the variance of the individual-specific scales and the (residual) preferences heterogeneity respectively (Equation 7) (Fiebig et al. 2010b).

$$\beta_{nk} = \tau_n \mu_k + [\gamma + \tau_n(1-\gamma)] L\sigma_{nk} \qquad (7)$$

Under this specification, both the MMNL and MNL models can be seen as special cases of the more general GMNL model. The GMNL model collapses to the MMNL model when $\tau_n = 1\ \forall(n)$, and to the MNL model when both $\tau_n = 1$ and $\sigma_{nk} = 0\ \forall(n)$. Because of this flexible specification of the indirect utility function, the log-likelihood function is no longer

concave and requires simulation procedure (i.e. simulated log-likelihood). We decreased the risk of local solution and "false" convergence by using 100 sets of starting values (built upon initial MNL estimates) and specified 500 Halton draws (r) for the simulation procedure. One limitation of the MMNL and GMNL models is that distributions of preferences over the sample need to be specified a priori. We tested different types of distributions (normal, log-normal, truncated normal) to determine which one was the most appropriate. The different models were estimated separately for the BWS and the DCE methods and we used Bayesian information criterion (BIC) to compare models performance.

Whilst it is mathematically impossible to separate variability in "tastes" from variability in "scales" (Hess and Rose 2012), it should be possible to account for both types of variability using a highly flexible mixed MNL model with a full covariance matrix (with variability in "scales" being captured by the variance-covariance elements). However specification of such a MMNL model would require estimating a very large number of parameters. In our study, we would need to estimate 18 mean parameters, 18 variance parameters and 153 covariance parameters. In this context the GMNL model is seen as a parsimonious version of the fully flexible MMNL model. As such the GMNL model could be eventually seen as a compromise between a "*not flexible enough*" and "*too flexible*" MMNL models.

Appendix 3: Rescaling procedure for estimates preferences

- *Step 1* (for BWS): Rescale the BWS estimates so that effects can be interpreted relative to the same reference levels as for the DCE. Assuming that the BWS estimates are located on a ratio scale, for any level (l) of the attribute (k) initially estimated relative to a reference attribute level (A1), it is possible to change the reference level by taking the difference with the estimate of interest (e.g., A3).

$$\hat{\beta}(BWS) = \{\hat{\beta}_{A2}^{A1}, \dots, \hat{\beta}_{A(L_A)}^{A1}, \hat{\beta}_{B1}^{A1}, \dots, \hat{\beta}_{B(L_B)}^{A1}, \dots, \hat{\beta}_{K(L_K)}^{A1}\}$$
$$Rescaling: \hat{\beta}_{k(l)}^{A1} - \hat{\beta}_{k1}^{A1} = \hat{\beta}_{k(l)}^{k1}, \forall\ k \neq A\ and\ \forall\ l \neq 1$$

- *Step 2* (for both BWS and DCE): Locate all the estimates on the same scale by dividing them by a common denominator (metric). Assuming between-attribute homoscedasticity (i.e., errors do not depend on the type of attributes considered), the ratio of two attribute levels estimates ($\hat{\beta}$) no longer depends on the scale parameter (σ_ε).

$$Rescaling: \frac{\frac{\hat{\beta}(DCE,BWS)_{k(l)}^{k1}}{\sigma_{\varepsilon(ntk)}}}{\frac{\hat{\beta}(DCE,BWS)_{M2}^{M1}}{\sigma_{\varepsilon(ntM)}}} = \frac{\hat{\beta}(DCE,BWS)_{k(l)}^{k1}}{\hat{\beta}(DCE,BWS)_{M2}^{M1}}, \forall\ k \neq M\ and\ \forall\ l \neq 1$$

- *Step 3* (for both BWS and DCE): Consider only the absolute values of the estimates, because the estimates from the DCE and BWS methods differ in their meaning. (DCE estimates indicate the influence of the attribute levels on the probability of choosing a health state profile whilst BWS estimates indicate the influence of the attribute levels on the probability of selecting a particular component of the health profile).

$$Rescaling: \frac{\hat{\beta}(DCE,BWS)_{k(l)}^{k1}}{\hat{\beta}(DCE,BWS)_{M2}^{M1}} = \left|\frac{\hat{\beta}(DCE,BWS)_{k(l)}^{k1}}{\hat{\beta}(DCE,BWS)_{M2}^{M1}}\right| \quad \forall\ k \neq M\ and\ \forall\ l \neq 1$$

Appendix 4: Monotonicity testing for the BWS method

Given two bundles of commodities $\{C_1,C_2\}$, such as health states defined along different QoL dimensions, preferences are monotone if the utility function (U) is non-decreasing; that is when $C_1 \geq C_2$, $U(C_1) \geq U(C_2)$ (Mas-Colell 1995). Whilst this definition can be directly applied to observed DCE choices, adjustment is required for BWS responses. Given BW responses were analysed within the random utility maximisation (RUM) framework; observed choices were interpreted as a result of comparing utilities of different commodities (attribute levels instead of alternatives). When modelling the probability of an item (attribute level) being selected as either BEST or WORST, the utility score is expected to reflect both the item importance (IMP) and its valence (VAL). It then becomes possible to extend the traditional micro-economic definition of monotonicity to BW choices by making some additional assumptions about the nature of the BW choices:

- **H1**: All different commodities (attribute's levels) belong to the same set of commodities (i.e. $\{C_1,C_2\} \subset C$) and are comparable. Given the attributes and levels in our study all describe consequences of glaucoma on individuals' quality of life and participants were individuals with glaucoma, then the attributes and level could be considered comparable.
- **H2**: An item is more likely to be selected either as BEST or WORST when it is perceived as being more important.
- **H3**: The level of severity defines the item's valence. For example, a low level of severity should be positively valued and conversely a high level of severity should be negatively valued.

Following the literature on QoL instruments, an important issue is to determine how importance (IMP) and valence (VAL) should be combined to obtain a utility score (Trauer and Mackinnon 2001; Hsieh 2012). A Multiplicative function is widely used: $V' = \alpha_1 IMP + \alpha_2 VAL + \beta(IMP \times VAL)$. We use the same approach to derive monotonic conditions for BW choices that can be empirically tested in particular tasks, such as tasks in which only one item is at the lowest severity level and all the other items at the highest severity level. In our study we cannot separately identify (α_1) from (α_2), but for any $\alpha_1 \in R$, $\alpha_2 \in R^+$ and $\beta > \alpha_1$ the following monotonic conditions hold:

- **V'(High importance; Low severity) > V'(Low importance; Low severity)**, such that a highly important & not severe item should be preferred over a not important & not severe one (i.e., item should be selected as BEST)
- **V'(High importance; Low severity) > V'(Low importance; High severity)**, such that a highly important & not severe item should be preferred over a not important & severe one
- **V'(Low importance; Low severity) > V'(Low importance; High severity)**, such that a not important & not severe item should be preferred over a not important & severe one

In the below illustration one can see that an item that is both "important" (scale = 1) and "low severity" (scale = 1) has a rescaled value score of 1. In comparison, rescaled scores for "not

important" & "low severity", "not important" & "high severity", and "important" & "high severity" are 0.5, 0.17 and 0. This ordering of the scores indicates that "important" & "low severity" items should be selected as BEST (or at least more likely to be selected as such) and at the opposite, "important" and "high severity" items should be selected as WORST.

Table. Illustration of the utility (V') score when α1 = α2 = 0.5; β = 1
Value function: V' = α1.IMPORTANCE + α2.SEVERITY + β.(IMPORTANCE.SEVERITY)

	(scale)											
Important	1	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00
	0.9	0.95	0.86	0.76	0.67	0.58	0.48	0.39	0.30	0.20	0.11	0.02
	0.8	0.90	0.81	0.73	0.64	0.55	0.47	0.38	0.29	0.21	0.12	0.03
	0.7	0.85	0.77	0.69	0.61	0.53	0.45	0.37	0.29	0.21	0.13	0.05
	0.6	0.80	0.73	0.65	0.58	0.51	0.43	0.36	0.29	0.21	0.14	0.07
	0.5	0.75	0.68	0.62	0.55	0.48	0.42	0.35	0.28	0.22	0.15	0.08
	0.4	0.70	0.64	0.58	0.52	0.46	0.40	0.34	0.28	0.22	0.16	0.10
	0.3	0.65	0.60	0.54	0.49	0.44	0.38	0.33	0.28	0.22	0.17	0.12
	0.2	0.60	0.55	0.51	0.46	0.41	0.37	0.32	0.27	0.23	0.18	0.13
	0.1	0.55	0.51	0.47	0.43	0.39	0.35	0.31	0.27	0.23	0.19	0.15
Not important	0	0.50	0.47	0.43	0.40	0.37	0.33	0.30	0.27	0.23	0.20	0.17
	(scale =>)	1	0.8	0.6	0.4	0.2	0	-0.2	-0.4	-0.6	-0.8	-1
		Low severity										**High severity**

Remark: Initial value scores are rescaled between 0 and 1 to ease their reading

Appendix 5: Investigating the impact of order randomisation on the preferences measurement

In our study the order of the attributes' levels was randomised across the tasks and within the participants meaning that each participant faced different orderings of the multi-attribute information. The order randomisation was done by ordering the attributes by levels of severity (from "No" to "Severe" difficulty; see Figure 1 from the main article). As noted by one of the reviewers, this pseudo-randomisation might have unintended consequences on the participants' choices as it prevents them becoming familiar with the content of the choice experiment (i.e., they cannot build expectations about where to find particular types of information).

Expected consequences for the BWS (case 2) method

As the method only requires participants to make within-profiles comparisons, the pseudo-random ordering of the items (attributes' levels) is not expected affect the measurement of preferences. It may be a desirable feature because it minimises the impact of ordering effects on the BWS results. For example, an item being located at the top or bottom of the profile may make it more like to be chosen as best or worst. This is akin to randomising the order of profiles within a DCE task.

To the best of our knowledge our study is the first to investigate ordering biases in BW decisions, and our results show that bottom-located items are significantly less likely to be selected either as BEST or WORST than top-located items. Our result suggests that future BWS (case 2) studies should be aware of ordering effects for the items within the profiles.

Expected consequences for the DCE (pairwise comparison) method

We acknowledge that pseudo-randomising the order of the attributes' levels within the options is unconventional and may adversely affect the quality of the participants' choices. However, there is not previous empirical research on this issue. The detrimental effects could arise because the changing order of attributes increases the subjective complexity of the choice tasks and/or promotes the use of heuristics that are not captured by standard random utility modelling of the choices.

As our study was not specifically designed to investigate the effect of randomising the order of the attributes' levels within the choice options, it is impossible to directly observe changes in participants' preferences (or more generally in data quality). However we tried to tackle this issue with an indirect approach looking at learning and fatigue effects in participants' choices. We estimated a scaled MNL model (also known as heteroscedastic MNL model) allowing the error variance (taken as a proxy measure for choice consistency) to dynamically change over the sequence of choice tasks. Previous DCE studies have found a significant learning effect in early choice sets (improved consistency) that levels off as respondents become experienced in the task (Caussade et al. 2005; DeShazo and Fermo 2002). In our study, if within-option order randomisation was associated with a large increase in the cognitive

burden of the choice tasks, then we expected these dynamic changes in choice consistency to be suppressed. For example, the participants' choices would remain largely random (i.e., equally inconsistent) throughout the sequence of tasks. We first divided the sequence of 36 tasks into 6 blocks of 6 tasks each (i.e., block #1-#6; block #7-#12; etc.). Then we used these different blocks as predictors of systematic changes in level of choice consistency (Block #1-#6 being used as reference point).

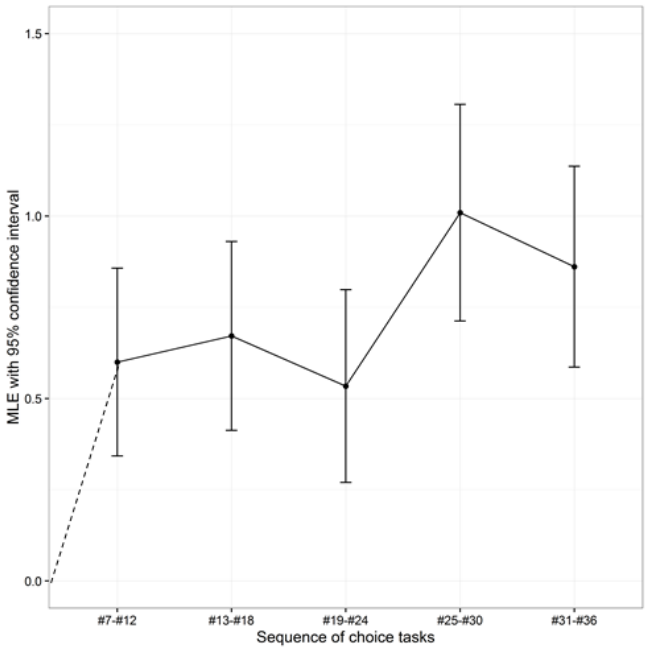
$$\lambda_{ntj} = exp\big(\alpha_1 BLOCK\_2_{ntj} + \alpha_2 BLOCK\_3_{ntj} + \alpha_3 BLOCK\_4_{ntj} + \alpha_4 BLOCK\_5_{ntj} + \alpha_5 BLOCK\_6_{ntj}\big)$$

$$V_{ntj} = exp\left(\lambda_{ntj}\big(\beta' X_{ntj}\big)\right)$$

In line with hypothesis of asymmetrical learning & fatigue effects we expect to find the following results:

- **H1:** $\boldsymbol{\alpha_k > 0, \forall k = \{1, 2, 3, 4, 5\}}$ – This represents the asymmetrical "learning" effect (i.e., individuals' choices become more consistent as they progress through the tasks sequence) [It is asymmetrical because (α_5) is also expected to be positive]
- **H2:** $\boldsymbol{\alpha_1 = \alpha_2 = \alpha_3}$ – This represents the period between learning and fatigue effects during which choice consistency is not expected to significantly change
- **H3:** $\boldsymbol{\alpha_4 > \alpha_5}$ – This represents the "fatigue" effect (i.e., individuals' choices become less consistent towards the end of tasks sequence)

Figure. Change in errors variance (choice consistency) over the sequence of choice tasks



The reference point is the beginning of the sequence of tasks (i.e., Tasks #1-#6).

All remaining effects are positive and significant, indicating thus that participants make more consistent choices in later tasks, what is usually interpreted as evidence for a learning effect as participants form their preferences and become familiar with the format of the choice tasks). Our **H1** hypothesis is supported. After forming their preferences and learning about the task, the consistency of individuals' choices should remain stable. Our **H2** hypothesis is verified as coefficients associated to the middle tasks are not significantly different from each other (α_1 = 0.600; α_2 = 0.672; α_3 = 0.534). Our **H3** hypothesis corresponds to the existence of a "fatigue" effect but is not supported by our data. Choices made at the end of the tasks sequence appear to be even more consistent than middle choices (However differences don't reach significance at 95% confidence level). This result could indicate a shift in respondents' decision rules. As a consequence of a fatigue effect, participants might have adopted "simplifying" decision rules that then make their choices more consistent (e.g., *Always choosing the alternative with the lowest level of a particular attribute*). We also examined how the choice proportions evolve over time (Block #1: 53.2-46.8; Block #2: 42.8-57.2; Block #3: 54.5-45.5; Block #4: 39.4-60.6; Block #5: 26.9-73.1; Block #6: 30.7-69.3), and overall it indicates that participants' choices were not completely clustered on one option or the other.

Our results show that despite the randomisation of attribute order respondents learn during which time their responses become more consistent and then settle at a level of response consistency. This is in line with response patterns in DCEs without attribute order randomisation. We acknowledge that without the counter-factual of a fixed attribute order DCE in our context we cannot definitively state that randomising the attribute order did not affect the DCE response quality.

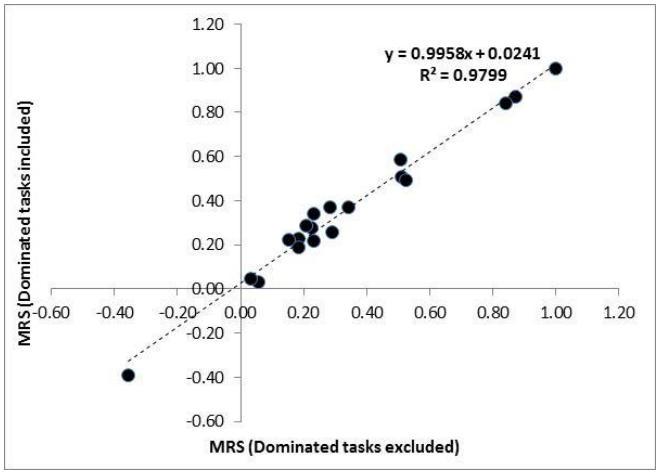
Appendix 6: Impact of including/removing dominated choice tasks on the preferences measurement

In our study we used five dominated tasks for the DCE method to investigate the monotonicity of respondents' choices (i.e., to prefer less severe health states over more severe ones). In dominated tasks, one option (health state) was clearly better than the other one on every aspect. As noted by one of the reviewers, including information from dominated tasks could bias preference estimates for two reasons:

- **Imbalance in misleading information across the attributes –** This would happen when a suboptimal decision (i.e., choosing the dominated option) generates more misleading information about preferences for some attributes (e.g., preferring "high severity" over "no severity") than others (e.g., preferring "mild severity" over "no severity").
- **Unaccounted for dominance-led heteroscedasticity –** The choices made in dominated tasks are expected to be more consistent than in non-dominated ones, because dominated tasks are presumably easier to answer.

Regarding the BWS (case 2) method, the notion of dominated tasks is itself open to discussion (see Appendix 4) and so far we know our study is the first to investigate this property for the BWS method. Regarding the DCE method, we do not expect our estimated to be biased because (1) the dominated tasks represent a small share of the total number of tasks (14 % of the DCE tasks are dominated), (2) overall participants performed well (73% of respondents made the "right" choice in all five dominated DCE tasks), and (3) the content of the dominated tasks systematically varied from one task to the other. We formally address this potential issue by comparing marginal rates of substitution (MRS) obtained from the MNL model estimated before/after having excluded information from dominated tasks. Overall the results indicate a very good level of matching between the two sets of estimates (R^2 = 98%), suggesting thus that the decision to include/exclude dominated tasks from the analysis has very little impact on the final results (see below figure).

Figure. Comparison of marginal rates of substitution (MRS) between the MNL models including vs. excluding information from dominated tasks



Note: Marginal rates of substitution (MRS) are obtained by dividing all MNL estimates by the estimate for "Vision: Severe" which appears to be the most influential attribute level in both MNL models.