HESG paper - A078

Multi-attribute valuation: A comparison of the discrete choice experiment and best-worst scaling approaches

Nicolas KRUCIEN, Verity WATSON, Mandy RYAN
Health Economics Research Unit, University of Aberdeen

Health Economists' Study Group (HESG) meeting, 8-10 January 2014, Sheffield, UK

Corresponding author: Dr Nicolas Krucien, nicolas.krucien@abdn.ac.uk, Health Economics Research Unit (HERU) - University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD, Scotland

1. Introduction

During the last fifteen years the attributes-based discrete choice experiments (DCEs) approach has been increasingly used to investigate preferences in a variety of health areas (Ryan & Gerard, 2003; de Bekker-Grob et al, 2012). More recently, another attributes-based approach, Best-Worst Scaling (BWS) (Louviere & Woodworth, 1990), has been introduced to health economics (Flynn et al, 2007) to study individual preferences (see Appendix for summary of applications of BWS in Health). A number of studies have compared these approaches in health, providing mixed evidence. Flynn et al (2011) and Potoglou et al (2011) compared the DCE [binary/pairwise choice] and BWS [case II] methods and showed similar results. However, Whitty et al (2013) performed the same comparison and found large discrepancies between the two approaches.

This current study adds to this limited literature by comparing variants of the BWS and DCE approaches in 2 areas: (i) elicitation of patients' preferences for attributes of a general practice service and (ii) elicitation of patients' utility weights for a health related quality of life index for glaucoma. Section 2 presents an overview of the two approaches and reflects on the advantages and limitations of BWS compared to DCEs; Section 3 describes the datasets, and presents the analytical methods; Section 4 presents the results and Section 5 provides a discussion of the results and a conclusion for this study.

2. Overview of DCEs and BWS

The DCE and BWS approaches have three factors in common: they are both (i) *survey-based* methods that ask respondents to make (ii) *hypothetical decisions* about (iii) *multi-attributes alternatives.*

2.1 Discrete Choice Experiment

The DCE approach presents respondents with choice sets containing multi-attribute descriptions of the good being valued and asks them to choose their preferred alternative (i.e. preferential choice). Each alternative in the choice sets is described by different levels of

attributes. Participants' observed choices are assumed to provide information on their underlying preferences for the different attributes of the good (Ryan et al, 2008). There are several variants of the DCE question that differ in the number of alternatives presented in a choice set and whether or not the choice set contains an opt-out alternative. The most used variants in health are the DCE binary choice (e.g. *would you accept this alternative? Yes/No*) and the DCE pairwise choice, hereafter referred as 'standard' DCE approach (e.g. *which alternative do you prefer? A/B*). An opt-out option is usually added to this choice set to allow for non-demanders.

2.2 Best Worst Scaling

Since its first application in food safety (Finn & Louviere, 1992), three variants of the BWS approach have been developed:

BWS case I (or object-case) asks respondents to make their best and worst decisions from a set of attributes. The observed decisions are used to measure the location of different attributes on the same underlying scale (e.g. importance, utility). This variant was developed as an alternative to the rating scale (rather than DCEs) to force respondents to discriminate between attributes, thus avoiding multiple attributes being valued equally. However, the DCE approach can also be used to valuing items on the same latent scale (Lancsar et al, 2007).

BWS case II (or profile-case) asks respondents to select the best and worst attribute levels within each alternative. For example, in their study on patients' preferences for dermatology consultation, Coast et al (2006) asked participants to select the best and worst attributes' levels of each alternative as well as to make a preferential choice. In this context, the BWS case II approach can be seen as a *modified*-version of the DCE binary choice where respondents first make within-alternative choices (i.e. *which attributes' levels is the best & worst?*) and second between-alternative choices (i.e. *is the proposed alternative better than doing nothing/current situation?*). In practice the BWS case II approach is used in addition to the DCE approach to obtain more information on the individuals' preferences, allowing the researcher to model the preferences for all but one attributes' levels on the same scale.

BWS case III or (multi-profile-case) asks participants to make their best and worst decisions from a subset of alternatives. This variant can be seen as an *extended*-version of the DCE pairwise choice where participants are asked to perform multiple ordering-choices (i.e. *which is the best alternative, the worst, the 2nd best, etc*) rather than one choice (*which is the best or worst*). In practice, the BWS case III approach is used to collect more information on individuals' preferences, allowing the researcher to model the preferences of each respondent separately.

2.3 A reflection on advantages and limitations of BWS compared to DCEs

Given the increased use of the BWS approach in health, in this section we reflect on its main advantages and limitations compared to the 'standard' DCE approach.

2.3.1 Advantages of BWS compared to DCEs - By asking participants to perform multiple-choices within the same task, BWS cases II and III collect more information on individuals' preferences than the DCE approach. This additional information can be used to obtain more precise estimates of the individuals' preferences. In turn, for a given level of precision in the preferences' estimates, the BWS approach can be used either to decrease the number of tasks per participant or the number of participants. The additional information can also be used to model individual preferences or to estimate preferences for more attributes' levels (since BWS estimates for discrete attributes are relative to one reference level over the attributes whereas DCEs estimates are relative to a reference category for each discrete attribute).

The BWS case II is easier to design than the 'standard' DCE approach. Indeed, the design for a BWS case II experiment can be obtained directly from an orthogonal main effects plan (OMEP) ensuring level balance and orthogonality between the attributes' levels (with no requirement to generate choice sets).

It has been argued that the BWS task is easier to complete than the DCE because it takes advantage of the respondents' propensity to best identify 'objects' located at extremities of a measurement scale (e.g. utility). Respondents to BWS tasks are therefore expected to be more consistent in their best and worst decisions rather than in their preferential choices.

Data collected with the BWS case I approach are easier to analyse because they do not require the use of statistical modelling techniques such as discrete choice models (DCMs). That is, the count approach based on the number of times each attribute is selected as best and worst leads to results which are closely related to those obtained with more sophisticated statistical techniques. This non-parametric analytical approach has the advantage of analyzing the observed decisions without making questionable assumptions about the decision-making processes used by the respondents.

Finally, the BWS approach allows the investigation of other psychological dimensions of the individuals' decision-making (other than preferences). The BWS approach is based on best and worst decisions given a particular (latent) scale or psychological construct which needs to be specified *a priori* by the researcher. For example, one can use an importance scale (e.g. *which are the most and least important items?*) or an attractiveness scale (e.g. *which are the most and least desirable items?*). Unlike the 'standard' DCE approach which is used to obtain preferential information, the BWS approach can be used in broader context, for example in satisfaction studies or quality analysis.

2.3.2 Limitations of BWS compared to DCEs - Depending on the study objectives and context, some of the advantages associated with the BWS approach may also be seen as limitations.

Whilst the BWS cases II and III provide more information on the underlying preferences of respondents, there is currently no empirical evidence on the quality of this additional information. The determinants of the best and worst decisions can differ, thus pooling all the information might not be invalid. Thus it would be more appropriate to

analyse separately the information provided by the best and worst decisions. If the additional information is of low 'quality' (as measured by the inconsistencies in the observed decisions), then it might be seen as an additional source of (statistical) noise in the analysis of the individuals' preferences, thus decreasing the ability of the experiment to detect these preferences.

Because of its characteristics (i.e. no opt-out option, observed decisions not based on consumption/uptake choices, no reference to the weak axiom of revealed preference), the BWS approach cannot be viewed as an economic method. In the standard economic approach, only choice data are relevant for the analysis of individual welfare (i.e. alternative [x] is deemed better than alternative [y] if and only if the individual would choose [x] over [y]) (Gul & Pesendorfer, 2005). Thus, the ranking-type (non-choice) data provided by the BWS approach are not in line with economics. In practice, this limitation is attenuated by assuming respondents are demanders of the proposed attributes/alternatives and would choose the "best" attribute/alternative when given the opportunity. This limitation can also be overcome by asking the participants to perform a DCE task in addition to the BWS task. However the interplay between the BWS and DCE approaches within the same tasks remains an open issue. Asking respondents first to perform their best and worst decisions may have an influence on their subsequent consumption choice, for example by favouring the use of a particular heuristic to assess the utility of each alternative.

Regarding the theoretical foundations of the two approaches, the DCE method is well-grounded in the consumer theory (McFadden, 1999) and consequently the observed (discrete) choices can been tested against theoretical axioms (e.g. transitivity, monotonicity) (Hougaard et al, 2012; McIntosh & Ryan, 2002; San Miguel et al, 2005; Ryan et al, 2009). However, it is not clear if the BWS method fulfils the same theoretical requirements, nor how to test the underlying axioms. More generally, the cognitive or psychological processes underlying the best and worst choices remain largely unknown and might be significantly different from those underlying DCE tasks. For example, with the BWS case II approach, respondents may struggle to select the worst attribute level when all the attributes are set to their best level and/or to select the best attribute level when all the attributes are set to their worst level (Al-Janabi et al, 2011).

Whilst BWS allows estimating the individuals' preferences for more attributes' levels than the 'standard' DCE approach, it is not clear if the BWS approach allows meaningful estimation of the preferences for a 'metric' attribute such cost in order to derive willingness-to-pay or accept (WTP/A) measures. Such a metric is usually used in the DCEs and is required for welfare analysis.

Despite claims that the BWS approach would be easier for the participants than the 'standard' DCE approach, there is a lack of empirical evidence to support this. Some studies have shown that the BWS approach might be more difficult than the DCE, at least for some research areas. For example, Whitty et al (2013) showed that participants to both a DCE [pairwise choice] and BWS [case II] experiments perceived the DCE tasks as being significantly easier than the BWS with 72% preferring the DCE tasks over the BWS ones.

Finally, the realism of the BWS format is questionable - participants are forced to make discriminatory decisions by identifying only one attribute as best and another one as worst. However in practice it is likely that some attributes are equally valued by the participants and thus a possibility of *ex aequo* should be introduced in the BWS tasks. A direct consequence of ruling out this possibility is that the final ranking of the attributes heavily depends on the experimental design (i.e. items frequency). For example, assuming that the participants to a BWS experiment consider several attributes as being of similar importance, the observed best (or worst) decisions among these attributes are expected to be randomly made. If the smallest possible design was used for that experiment, where each attribute was paired only once with any other attribute, then the random decisions are likely to result in unequal frequencies of choices (as best or worst) for the attributes of the same importance. In turn, these artificial differences in choice frequencies between the attributes result in different ranks of importance.

Given these potential advantages and limitations of the BWS approach over the DCE approach, it remains unclear if the former should be considered as an *improved*-version of the 'standard' DCE approach leading thus to more precise but qualitatively equivalent results. Alternatively, the BWS approach can be seen as a new and independent research method with the potential for bringing additional insights in the individuals' decision-making with qualitatively different results. In this study this issue is investigated by performing two BWS/DCE comparisons - we compare the BWS case I and DCE binary choice approaches applied to elicit patients' preferences for a general practice care (GP study) and the BWS case II and DCE pairwise choice approaches applied to elicit patients' utility weight for a Glaucoma condition specific utility index (Glaucoma study).

3. Empirical work comparing BWS and DCEs

3.1 Presentation of the case studies

3.1.1 GP study - This study elicited preferences of patients suffering from multiple chronic conditions for changes in general practice care as recommended by the Chronic Care Model (CCM) (Wagner et al, 2001). The study took place in France and used a within-subjects design in which 150 patients completed both the BWS case I and DCE binary choice tasks. Further information on recruitment and sampling strategy are available in Krucien et al (*in press*)[1].

The BWS case I tasks used 10 attributes corresponding to different recommendations of the CCM to improve general practice care[2]. These attributes were combined in a balanced incomplete block design (BIBD) resulting in 6 BWS tasks (An example is presented in Figure

[1] The study was approved by a local ethic committee (*Comité de protection des personnes "Ile-de-France VI", Paris, France*). Participation in the study was voluntary and respondents could end participation at any time.

[2] [Communication] GP can be reached by phone or email; [Empowerment] GP helps the patients to manage their chronic conditions themselves; [Habits] GP pays attention to the patients' health habits; [Information] GP is in touch with other professionals involved in the patients' chronic care; [Monitoring] GP checks regularly the appropriateness of the patients' chronic care; [Nurse] GP works with a specialised nurse or other paramedical professional to provide the chronic care; [Planning] GP organizes the patients' chronic care on the long-term; [Coordination] GP coordinates the different medical services related to patients' chronic care; [Decision] GP involves the patients in the medical decision making; [Support] GP takes into account the psycho-social dimension of the chronic conditions

1.A). Each task included five attributes and each attribute appeared three times over the tasks. The participants were asked to select the most and least important attributes.

Figure 1.A. Illustration of the BWS tasks used for the GP study

Attributes	Most important	Least important
The general practitioner manages the different medical services required to treat your chronic conditions.	✘	
The general practitioner involves a nurse or other paramedical professional to improve your chronic care.		
The general practitioner asks questions about your health habits and gives you advices to improve your habits.		
The general practitioner organizes your treatment on the long-term, for example s/he informs you of the annual number of medical consultations and which health professional you should to contact in case of need.		✘
The general practitioner contacts the other health professional(s) you recently consulted to stay informed of your medical situations.		

The DCE tasks used the same 10 attributes, each described with two extreme levels (*Achieved, Not achieved*). The DCE tasks were designed according to an attributes-blocked design (Witt et al, 2009), allocating attributes in two sets of six-specific attributes and keeping two-common attributes across the sets to ensure comparability. For each set, an OMEP design provided eight choice tasks. Participants were randomly allocated to one of two versions of the DCE questionnaire; for each of the eight tasks they were asked to refuse or to accept the proposed alternative (An example is presented in Figure 1.B).

3.1.2 Glaucoma study - This study developed a Glaucoma utility measure. The study took place in England and used a within-subjects design in which 293 patients completed the BWS case II and DCE pairwise choice tasks. Further information on the recruitment and sampling strategy are available in Burr et al (2007)[3].

In both the BWS and DCE tasks, six attributes are used to describe a Glaucoma related health state[4]. Each attribute is described with the same four levels of increasing impairment (i.e. *No difficulty; Some difficulty; Quite a lot of difficulty; Severe difficulty*). The BWS tasks were selected using an OMEP resulting in 32 sets. The participants were asked to select the best and worst attribute's levels in each of the 32 BWS tasks (An example is presented in Figure 2.A).

The DCE tasks used the same six attributes and corresponding four levels. The choice tasks were designed by taking the profiles presented in the BWS tasks as alternative A in the choice set and pairing this with its fold-over as alternative B. Thus there were also 32 choice sets and the respondents were asked to select worse health state. (An example of DCE task is presented in Figure 2.B).

[3] Ethical approval was obtained for each phase of the study from the Central Office of Research Ethics Committees. The research was conducted according to the tenets of the Declaration of Helsinki.

[4] [Vision] Central and near vision; [Light] Lighting and glare; [Mobility] Mobility; [Daily] Activities of daily living; [Eye] Local eye discomfort; [Other] Other effects of Glaucoma and its treatment

Figure 1.B. Illustration of the DCE tasks used for the GP study

Description of the chronic care		
The general practitioner **does not manage** the different medical services required to treat your chronic conditions.		
The general practitioner **involves** a nurse or other paramedical professional to improve your chronic care.		
The general practitioner **does not ask questions** about your health habits and gives you advices to improve your habits.		
The general practitioner **organizes** your treatment on the long-term, for example s/he plans all the follow-up visits for a year and explains you what to do in case of need.		
The general practitioner **does not contact** the other health professional(s) you recently consulted to stay informed of your medical situations.		
Would you accept this medical care to manage your chronic conditions? (Pick only one box)	**No** ☒	**Yes** ☐

Figure 2.A. Illustration of the BWS task used for the Glaucoma study

Best aspect	**Aspects of situation A**	*Worst aspect*
	Some difficulty with activities of daily living	
	Some difficulty with local eye discomfort	
✘	**Some difficulty** with the effects of glaucoma and its treatments	
	Quite a lot of difficulty with lighting and glare	
	Severe difficulty with central and near vision	✘
	Severe difficulty with mobility	

3.2 Empirical analysis of DCE and BWS data

3.2.1 Analysis of DCE data - For the GP study, the DCE data were analysed using a binary logit model, including an additional individual error component to account for correlation of observations within the respondents. A dummy coding strategy was used, with the 'theoretically' worst level omitted (i.e. *No achievement*).[5]

For the Glaucoma study, the DCE data were analysed using a conditional logit model, including an additional individual error component to account for potential correlation of the observations within the respondents (i.e. error component logit model). A dummy coding strategy was also used, with the 'theoretically' best level omitted (i.e. *No difficulty*).[6]

[5] The model therefore included 13 parameters to estimate (1 parameter per attribute x 10 attributes + 2 alternative-specific mean and SD parameters + 1 version-specific constant parameter [V1 vs V2]).
[6] The model included 20 parameters to estimate (3 parameters per attribute x 6 attributes + 2 alternative constant mean and SD parameters).

Figure 2.B. Illustration of the DCE task used for the Glaucoma study

Each choice question describes two health state situations: Situation A and B. Imagine that you have these difficulties and pick the scenario you think is **WORSE**. You may not like either situation but choose the one that is less preferable to you by putting a tick in the appropriate box. Lease tick just ONE box for every question.

SITUATION A	SITUATION B
No difficulty with: ● Central and near vision ● Lighting and glare ● Mobility **Some difficulty** with: ● Activities of daily living ● Eye discomfort ● Other effects of glaucoma and its treatment	**No difficulty** with: ● Central and near vision **Some difficulty** with: ● Lighting and glare **Quite a lot of difficulty** with: ● Activities of daily living ● Other effects of glaucoma and its treatment **Severe difficulty** with: ● Mobility ● Eye discomfort
(Tick only one box) ☐ **Situation A**	☒ **Situation B**

3.2.1 Analysis of BWS data - Analysis of data from BWS tasks has two approaches, one relying on a count analysis and the second drawing on the same analytical framework as DCEs (Flynn et al, 2007). We use the latter approach with a 'MaxDiff' specification of the choice model (i.e. modelling of the probability to select the pair of best and worst attributes maximising the underlying differences of utility/importance).

For the Glaucoma study, the BWS data were analysed using an error component MaxDiff logit model. The attribute level '*No difficulty in mobility*' was omitted, and its value set to 0. For the GP study, the BWS data were also analysed using an error component MaxDiff logit model and the attribute level '*Cooperation with a nurse*' was omitted. In both studies the reference was selected on the basis of a count analysis of the BW data which indicated that these attribute levels were least valued by respondents[7].

3.3 Comparison of BWS and DCE methods

This paper is opportunistically using two data sets that have both BWS and DCE data. Tests across studies therefore differ.

3.3.1 Feasibility - The feasibility of the BWS and DCE approaches was investigated in the GP study by asking respondents to report how difficult and interesting they found the BWS and DCE tasks. Both were measured on a 4-point rating scale. Descriptive statistics of responses was conducted. Observations were then merged in two categories ('*Difficult vs. Easy*' and

[7] For clarity of the presentation, this analysis is not reported here but can be obtained from the corresponding author on request.

'*Interesting vs. Not interesting*') and the equality of the distributions between the DCE and BWS tasks tested using the Chi-2 test of independence (α=5%).

3.3.2 Theoretical validity - Theoretical validity is defined in two ways: (i) ability to provide results consistent with *a priori* expectations and (ii) satisfaction of theoretical axioms underlying consumer theory.

3.3.2.1 A priori expectations - A priori expectations concern the sign and magnitude of the estimated preferences. For the BWS approach, for both case studies, all estimates are located on one-side of the measurement scale since the least important attribute level was selected as a reference level. Thus, all signs are expected to be positive. For the GP study, no *a priori* assumptions were formulated about the magnitude of the estimates. For the Glaucoma study, given the natural order of attribute levels, estimates were assumed to monotonically increase or decrease over the attributes' levels.

For the DCE, expected parameter signs depend on the task format. For the GP study respondents were asked to state if they considered the hypothetical alternatives as being acceptable, estimated preferences were therefore expected to be non-negative. No *a priori* assumptions were made regarding magnitude of preferences. For the Glaucoma study respondents were asked to indicate the worse alternative. Then any positive estimate would indicate a negative effect in terms of desirability. Given the wording of the attributes' levels, it was expected (i) to obtain strictly positive estimates and (ii) to find a monotonic increase in the estimates of a particular attribute.

3.3.2.2 Theoretical axioms - According to standard micro-economic theory, many axioms need to be satisfied in order to obtain well-behaved utility functions (Samuelson, 1938; Houthakker, 1950). This study compared the DCE and BWS approaches on two of these conditions:

Stability - The stability of preferences is defined as the ability of individuals to perform the same choices when confronted the same decision tasks. In the Glaucoma study, the stability condition was tested by repeating one DCE task and two BWS tasks. Preferences are considered stable if the task is completed the same way. For the BWS approach, the two repeated tasks provide four measures of choices stability: choice stability in the best choices and choice stability in the worst choices at the first task repetition and the second task repetition. Therefore, each respondent has a stability score for the BWS tasks ranging from "0" (i.e. no matching of the best and worst choices in the two repeated tasks) to "4" (i.e. full matching of choices). For purpose of comparison, the stability scores of the BWS tasks are re-coded in three categories: "no stability" (score=0), "low stability" (score=1) and "moderate to high stability" (score ≥ 2). The independence of the DCE and BWS distributions is tested with a Chi-2 statistic (α=5%).

Completeness - The completeness condition implies individuals have a set of well-defined preferences for the attributes' levels. The completeness condition was indirectly tested through the ability of the DCE and BWS responses to generate a preference ordering of attributes (Lagerkvist, 2013). This ordering is based on all pairwise attributes-comparison embedded within the choice tasks. Then the choices observed in the BWS task can be used to retrieve the preferences relations in the pairwise attributes-comparisons. Finally, the results of these comparisons are used to compute for each attribute (i) the probability of being preferred over each other attribute (j). Then this information can be used to rank order the attributes.

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

Where: (A) is the total of times attribute (i) is preferred over attribute (j); (B) is the total of ties in ranks and (C) is the total of times attribute (j) is preferred over attribute (i). For example, a BWS task including three attributes $\{X_1,X_2,X_3\}$ can be disaggregated into three pairwise attributes-comparisons $\{X_1X_2,X_1X_3,X_2X_3\}$. Assuming X_1 and X_3 were selected respectively as best and worst attributes in the corresponding choice task, one can infer the following preference relations in the pairwise attributes-comparisons: $X_1 \succ X_2$, $X_1 \succ X_3$ and $X_2 \succ X_3$. For two attributes $\{X_1X_2\}$, the repetition of these comparisons over the respondents and the choice tasks is then summarised as the number of times X_1 is preferred over X_2 [count A], X_2 is preferred over X_1 [count C] and indifference [count B]. If these counts are respectively of the following values {70,20,10}, then the probability of X_1 being preferred over X_2 is 72.7% and conversely the probability of X_2 being preferred over X_1 is 18.2%. These probabilities reflect the dominance relations between the attributes, which can be ranked accordingly. In this study, the relative performance of the BWS and DCE approaches was assessed using the Spearman (ρ) coefficient of correlation for the attributes rank-orderings. This coefficient varies between -1 and +1, with an absolute value close to 1 indicating high correlation between the rankings (Siegel & Castellan, 1988).

3.3.3 Convergent validity - Given that DCEs and BWS estimate the locations of several attributes (or attributes' levels) on a latent utility scale, convergent validity can be assessed by comparing the estimated locations between the two approaches. However, the parameter estimates obtained from the two approaches are not directly comparable because of differences in the anchors and latent scales (Potoglou et al, 2011). Thus, the estimates need to be rescaled before being compared.

3.3.3.1 Incomparability of the anchors - The DCE and BWS estimates differ in the anchor(s) or origin point(s) of the measurement scale. Whilst the BWS estimates locate all items on the same scale and the omitted item corresponds to the origin point of the scale, the DCE locates all but one item of each attribute on the same measurement with each omitted attribute level the origin point for each attribute. To overcome this BWS estimates are rescaled to have the same origin points as the DCE estimates. Initially, all BWS estimates are scaled relative to a same reference, say level '1' of attribute 'A' (A1). Assuming that the BWS estimates are

located on a ratio scale, the proposition below means that for any level (l) of the attribute (k) initially estimated relative to A1 ($\hat{\beta}_{k(l)}^{A1}$) it is possible to change the reference level by taking the difference with the corresponding estimate ($\hat{\beta}_{k1}^{A1}$). For the Glaucoma study, all the initial BWS estimates were rescaled relative to the 1st level (i.e. *No difficulty*) of each attribute. For the GP study, the initial BWS and DCE estimates already shared the same reference point and then there was not necessary to change the anchors the BWS estimates.

$$\hat{B}(BWS) = \{\hat{\beta}_{A2}^{A1}, \dots, \hat{\beta}_{A(L_A)}^{A1}, \hat{\beta}_{B1}^{A1}, \dots, \hat{\beta}_{B(L_B)}^{A1}, \dots, \hat{\beta}_{K(L_K)}^{A1}\}$$

$$\text{Proposition 1: } \hat{\beta}_{k(l)}^{A1} - \hat{\beta}_{k1}^{A1} = \hat{\beta}_{k(l)}^{k1}, \forall\, k \neq A \text{ and } \forall\, l \neq 1$$

3.3.3.2 Incomparability of the scales - The variance of the measurement scales may differ. Given that the estimates perfect confound the means and variances, in order to compare the estimates the variance parameter must be normalised. The scale parameter can be ruled out by expressing each estimate in the same units of measurement. Typically, this is accomplished by dividing each estimate by a metric (e.g. cost) with the resulting ratio providing a *free-scale* measure (e.g. willingness-to-pay). Under the assumption of between-attribute homoscedasticity (i.e. the errors that respondents make in their choices do not depend on the type of attributes considered), the proposition below indicates that the ratio of two attributes' levels estimates ($\hat{\beta}$) no longer depends on the scale parameter (σ_ε). For the Glaucoma study, all the DCE and rescaled BWS estimates were converted onto the same scale by dividing both by the same metric or denominator ('M') which was the 4th level (i.e. *severe difficulty*) of the 'Activities of daily living' attribute. For the GP study, the common denominator was the 'Informational continuity' attribute.

$$\text{Proposition 2: } \frac{\frac{\widehat{\beta}(DCE,BWS)_{k(l)}^{k1}}{\sigma_{\varepsilon(ntk)}}}{\frac{\widehat{\beta}(DCE,BWS)_{M2}^{M1}}{\sigma_{\varepsilon(ntM)}}} = \frac{\widehat{\beta}(DCE,BWS)_{k(l)}^{k1}}{\widehat{\beta}(DCE,BWS)_{M2}^{M1}}, \forall\, k \neq M \text{ and } \forall\, l \neq 1$$

For both case studies, we compare the BWS and DCE rescaled estimates in two ways. First, we regress the BWS estimates on the DCE estimates from the same study and the linear R^2 is used as measure of (linear) association between the two approaches. Second, we compare the rank-orderings of rescaled estimates using the Spearman (ρ) correlation coefficient (α=5%).

4. Results

4.1 Feasibility - For the GP study respondents perceived both the BWS case I and DCE binary choice tasks as easy and interesting to complete - 79.4% of respondents considered the BWS tasks as easy or very easy and 98% as interesting or very interesting. Regarding the DCE tasks, these proportions are 69.3% and 93.3% respectively for ease of completion and interest. The BWS tasks were significantly easier to complete than the DCE tasks (Chi-2=3.931, P=0.047, DF=1), but equally interesting (Chi-2=3.125, P=0.077, DF=1).

4.2 Theoretical validity

4.2.1 A priori expectations - For both DCEs, all estimated parameters have the expected signs (Columns 3 and 4 in Tables 1 and 2). However for the Glaucoma study, *a priori*

assumptions concerning the magnitude of estimates are not completely verified. For some attributes, the impact of an increase in the level of difficulties is higher for small increase rather than large increase. For example, for the "Light" attribute, the impact of having "Quite" difficulties (+0.36) is larger than having "Severe" difficulties (+0.27) which is about the same as "Some" difficulties (+0.26). Looking at the BWS estimates, there are also violations of *a priori* assumptions. Some attributes are associated with larger impact (importance) of intermediate levels (i.e. Some, Quite) than the extremes levels (i.e. No, Severe). There is especially an issue with the level of difficulty "Severe" which is consistently valued as being of lower importance than "Quite" difficulty. When comparing the two approaches, the DCE results are associated with 16% of violations in the *a priori* assumptions (3/18) and the BWS results with 44% of violations (16/36).

4.2.2 Theoretical conditions

Stability - In the Glaucoma study, 124 respondents provided sufficient data for the DCE stability test. The remaining respondents failed to complete either one or both of the repeated DCE tasks. Of these 124 respondents, 63.7% made the same choice in both choice tasks. Regarding the BWS, 243 respondents provided sufficient data for the stability test. The results indicated almost no stability in the individuals' BWS choices. Only 2 (0.8%) respondents out of 243 provided the same best and worst choices in the two BWS tasks.
Notably, respondents were more stable in their worst choices than in their best choices with 9 (3.7%) respondents providing the same best choices in the repeated tasks, whereas 166 (68.3%) respondents made the same choices for the worst attributes levels. The comparison of the BWS and DCE approaches was based on 188 respondents who provided sufficient data. The stability of responses to the DCE and BWS tasks is independent (Chi-2=0.259, DF=2, P-value=0.88), thus indicating that the stability of the choices made in one approach was not related to the stability of the choices made with the other approach.

Completeness - For the GP study, the analyses of the BWS case I and DCE binary choice data results in 10 and 7 ranks respectively out of 10 expected in case of perfect discrimination between the attributes' levels (Table 3). Looking at the 'Informational continuity' attribute, the BWS case I approach led to a 100% value thus indicating that this attribute was strictly preferred over all other attributes (i.e. in each pairwise attributes-based comparison the 'Informational continuity' was consistently preferred to any other attribute). With the DCE binary choice approach, the pairwise probability of the 'Informational continuity' decreases to 83%, thus indicating that this attribute is still a dominant attribute in the individuals' choices but it was not always selected as being the best one in all the pairwise attributes-based comparisons. Overall the results of the BWS approach are more evenly distributed along the 0-100 probability scale than the DCE results, thus suggesting that the choices observed with the BWS case I approach discriminated more between the attribute levels than those observed with the DCE approach. This result is in line with the objective of the BWS method to force the respondents to make discriminating choices. However, the Spearman coefficient of correlation is significantly different from 0 (ρ=0.825, P=0.003) and close to 1,

thus suggesting that the DCE and BWS approaches led to quite similar orderings of the attributes in terms of preferences.

Table 1. Regression results and convergent validity analysis for the Glaucoma study

Attribute	Level	Initial estimates β (p-value)		Rescaled estimates		
		BWS case II	DCE pairwise	BWS case II	DCE pairwise	BWS/DCE
Day	No	0.64 (0.001)	-	0.00	-	-
Day	Some	0.12 (0.03)	0.33 (0.001)	-0.51 *	0.35	-1.45
Day	Quite a lot	1.02 (0.001)	0.56 (0.001)	0.37	0.60	0.63
Day	Severe	1.66 (0.001)	0.94 (0.001)	1.00	1.00	1.00
Eye	No	0.99 (0.001)	-	0.00	-	-
Eye	Some	1.34 (0.001)	0.18 (0.001)	0.34	0.19	1.79
Eye	Quite a lot	1.71 (0.001)	0.16 (0.001)	0.71	0.17	4.15
Eye	Severe	0.98 (0.001)	0.24 (0.001)	-0.01	0.26	-0.04
Glaucoma	No	0.84 (0.001)	-	0.00	-	-
Glaucoma	Some	0.42 (0.001)	0.11 (0.008)	-0.41	0.12	-3.52
Glaucoma	Quite a lot	0.92 (0.001)	0.06 (0.222)	0.08	0.06	1.23
Glaucoma	Severe	0.13 (0.015)	0.22 (0.001)	-0.70	0.23	-2.97
Light	No	0.03 (0.512)	-	0.00	-	-
Light	Some	0.95 (0.001)	0.26 (0.001)	0.90	0.28	3.26
Light	Quite a lot	1.31 (0.001)	0.36 (0.001)	1.25	0.38	3.28
Light	Severe	1.06 (0.001)	0.27 (0.001)	1.01	0.29	3.52
Mobility	No	0	-	0.00	-	-
Mobility	Some	0.63 (0.001)	0.35 (0.001)	0.62	0.37	1.66
Mobility	Quite a lot	0.34 (0.001)	0.56 (0.001)	0.33	0.60	0.56
Mobility	Severe	0.88 (0.001)	0.97 (0.001)	0.86	1.03	0.84
Vision	No	3.1 (0.001)	-	0.00	-	-
Vision	Some	0.47 (0.001)	0.36 (0.001)	-2.58	0.38	-6.73
Vision	Quite a lot	1.16 (0.001)	0.62 (0.001)	-1.90	0.66	-2.88
Vision	Severe	0.9 (0.001)	1.17 (0.001)	-2.16	1.24	-1.73

* Explanations: The initial BWS estimates for the 'No', 'Some' and 'Severe' levels of the 'Day' attribute are respectively 0.64, 0.12 and 1.66. The 1st step of the rescaling process is to measure 'Day - Some' relative to 'Day - No' (as with the DCE approach) instead of 'Mobility - No'. Following the proposition 1, the estimate of 'Day - Some' relative to 'Day - No' is -0.52 (= 0.12-0.64). The rescaled BWS estimate for 'Day – Severe' is 1.02 (= 1.66-0.64). The 2nd step of the rescaling process is to locate all the estimates on the same scale by diving them with a common denominator. Following the proposition 2, the 'Day - Severe' estimate is used as common denominator and then the final rescaled BWS estimate for 'Day - Some' is -0.51 (= -0.52/1.02).

For the Glaucoma study, the analyses of the BWS case II and DCE pairwise choice data results in 15 (63%) and 13 (54%) ranks respectively for the attributes' levels out of 24 expected (Table 4). This result suggests that both approaches are only moderately able to discriminate between the attribute levels. In addition, the Spearman coefficient of correlation is not significantly different from 0 (ρ=0.051, P-value=0.81), thus suggesting that the two approaches led to different preferences orderings of the attributes' levels. Looking at the values (in %) for the 24 attributes' levels, 14 (58%) of them have relatively close values in the two methods (i.e. difference ≤ 20 points) and 10 (42%) have divergent values (i.e. difference > 20 points). Especially, the values provided by the two methods suggested a preference reversal for 4 attributes' levels (i.e. difference ≥ 50 points).

Table 2. Regression results and convergent validity analysis for the GP study

Attribute	Level (Realisation)	Initial estimates β (p-value)		Rescaled estimates		
		BWS case I	DCE binary	BWS case I	DCE binary	BWS/DCE
Informational continuity	Yes	3.44 (0.001)	1.26 (0.001)	1.00	1.00	1.00
Regular follow-up	Yes	2.14 (0.001)	0.82 (0.001)	0.62	0.66	0.94
Responsibility of the coordination	Yes	3.16 (0.001)	0.80 (0.001)	0.92	0.65	1.42
Psychological support	Yes	2.44 (0.001)	0.77 (0.001)	0.71	0.62	1.14
Advices on health habits	Yes	2.52 (0.001)	0.65 (0.001)	0.73	0.53	1.38
Shared decision making	Yes	1.77 (0.001)	0.60 (0.001)	0.51	0.49	1.05
Ability to reach at distance	Yes	1.85 (0.001)	0.52 (0.001)	0.54	0.43	1.25
Planned care	Yes	1.11 (0.001)	0.42 (0.001)	0.32	0.35	0.91
Empowerment of the patient	Yes	2.10 (0.001)	0.38 (0.002)	0.61	0.32	1.89
Collaboration with a nurse	Yes	0	-0.04 (0.697)	0.00	0.00	-

Table 3. Results of the completeness analysis for the GP study

Attribute	Superiority of the level in the binary comparisons with the other levels (in %)			Ranking of the levels in terms of superiority	
	BWS case I	DCE binary	Absolute diff.	BWS case I	DCE binary
Ability to reach at distance	33%	33%	0%	7	4
Advices on health habits	78%	33%	44%	3	4
Collaboration with a nurse	0%	0%	0%	10	7
Empowerment of the patient	44%	17%	28%	6	6
Informational continuity	100%	83%	17%	1	1
Planned care	11%	17%	6%	9	6
Psychological support	67%	67%	0%	4	2
Regular follow-up	56%	67%	11%	5	2
Responsibility of the coordination	89%	50%	39%	2	3
Shared decision making	22%	25%	3%	8	5

4.3 Convergent validity - For the GP study, the rescaled estimates for the BWS case I and DCE binary choice are presented in Table 1 (columns 5-7). The comparison showed similarities in the rescaled estimates and rankings provided by two approaches. Except for the 'Empowerment of the patient' attribute for which the rescaled BWS estimate was almost 2 times larger than the rescaled DCE estimate, all the other differences between the rescaled BWS and DCE estimates are quite small, thus indicating that the two approaches led to

qualitatively similar results in terms of individuals' preferences. The regression of the rescaled DCE estimates against the rescaled BWS estimates led to a high level of linear association (R^2=81.4%). The Spearman coefficient of correlation between the rankings was significantly different from 0 (P-value=0.003) and close to 1 (ρ=0.83).

For the Glaucoma study, the rescaled estimates for the BWS case II and DCE pairwise choice are presented in Table 2 (columns 5-7). In contrast to the above result, the comparison showed no discernible relationship between the rescaled estimates and rankings provided by the two methods. For several attributes' levels, the rescaled BWS and DCE estimates are very different in terms of magnitude, for example the rescaled BWS estimates for the 'Eye – Quite a lot' and 'Vision – Some' attributes' levels were respectively 4 and 7 times larger than the rescaled DCE estimates. In addition, the comparison of the two sets of rescaled estimates showed evidence of opposite results (i.e. preferences reversals) (cf. 'Vision – Some-attribute's level). These results suggest that the two approaches led to qualitatively different results. The regression of the rescaled BWS estimates against the rescaled DCE estimates resulted in a very low association (R^2=3.4%) and the Spearman coefficient of correlation between the rankings was not significantly different from 0 (ρ=0.032, P-value=0.9).

Table 4. Results of the completeness analysis for the Glaucoma study

Attribute	Level (Difficulty)	Superiority of the level in the binary comparisons with the other levels (in %)			Ranking of the levels in terms of superiority	
		BWS case II	DCE pairwise	Absolute diff.	BWS case II	DCE pairwise
Day	No	100%	40%	60%	1	8
Day	Some	45%	30%	15%	9	10
Day	Quite a lot	50%	60%	10%	8	5
Day	Severe	5%	70%	65%	14	3
Eye	No	55%	65%	10%	7	4
Eye	Some	45%	60%	15%	9	5
Eye	Quite a lot	55%	85%	30%	7	2
Eye	Severe	65%	65%	0%	5	4
Glaucoma	No	55%	45%	10%	7	7
Glaucoma	Some	50%	20%	30%	8	12
Glaucoma	Quite a lot	55%	50%	5%	7	6
Glaucoma	Severe	30%	15%	15%	12	13
Light	No	70%	30%	40%	4	10
Light	Some	35%	35%	0%	11	9
Light	Quite a lot	40%	65%	25%	10	4
Light	Severe	50%	60%	10%	8	5
Mobility	No	80%	30%	50%	3	10
Mobility	Some	45%	35%	10%	9	9
Mobility	Quite a lot	60%	25%	35%	6	11
Mobility	Severe	10%	50%	40%	13	6
Vision	No	90%	100%	10%	2	1
Vision	Some	45%	35%	10%	9	9
Vision	Quite a lot	65%	65%	0%	5	4
Vision	Severe	0%	60%	60%	15	5

5. Discussion and conclusion

This study compared variants of the BWS and DCE approaches, providing mixed evidence. The comparison between the BWS case I and DCE binary choice approaches highlighted

similarities in theoretical and convergent validity. However, BWS case I was able to marginally improve the results on the different criteria (i.e. less difficult, higher ability to discriminate individuals' preferences). These results suggest that to some extent the BWS case I and DCE binary choice might be seen as substitutes.

The comparisons between the BWS case II and DCE pairwise choices approaches indicated significant differences for all criteria, except the ability of the methods to rank order attributes. Our results showed a higher ability of the DCE approach to verify a priori assumptions and to satisfy the stability condition. The lack of convergent validity between BWS and DCE approaches suggests that the two approaches are not substitutes.

Previous studies comparing these approaches have provided contrasting evidence. Louviere & Islam (2008) also compared the BWS case I and DCE binary choice approaches and showed divergences between them. Flynn et al (2011) and Potoglu et al (2011) compared the BWS case II and DCE binary/pairwise choice approaches and showed similar results (R^2=81% in Flynn et al, 2011). However these studies are not directly comparable to ours because of differences in the experimental setup and/or the analytical approach. Louviere & Islam (2008) did not provide any quantification of the divergences between the DCE binary choice and BWS case I approaches. Flynn et al (2011) used the same tasks to collect both DCE binary choice and BWS case II data, potentially aligning DCE and BWS choices. In their comparison of the BWS case II and DCE pairwise choice approaches, Potoglu et al (2011) used an attributes-blocked design to generate the DCE task and in the comparison process only the DCE estimates were divided by a common metric (the BWS estimates were not completely rescaled). Recently, Whitty et al (2013) also compared the BWS case II and DCE pairwise choice approaches using an experimental design and analytical method close to ours. As in our study, Whitty et al (2013) showed important differences between the two approaches (Pearson correlation coefficient=0.286, p=0.283).

The comparisons in this study tried to complement existing studies. First, the BWS and DCE choices were collected using separate tasks. Second, comparability of the DCE and BWS tasks in terms of attributes' levels was carefully ensured, using a specific rescaling process. Third, additional criteria were assessed to extend comparisons (feasibility, theoretical validity).

Two arguments may help to explain our results: a consciousness issue and behavioural compatibility. Regarding the former, the BWS method is a 'direct' approach (respondents are explicitly asked to discriminate between the items according to the dimension of interest) and the DCE is an 'indirect' approach (with values inferred from observed choices). Whilst economists assume information processing across attributes/levels when completing valuation tasks, findings from consumer research suggest much decision-making is made unconsciously or mindlessly (i.e. the factors determining the decision making are not recognised by the individual making the decision). Dijksterhuis et al (2005) explains this with reference to shopping at the supermarket – an individual has often packed their shopping basket and is checking out with very little thought given to the importance of attributes of the goods purchased. A consequence of this is that when using direct preference elicitation methods, such as BWS, individuals often cannot accurately report the importance of a given attribute/attribute level. Mueller et al (2010) note that indistinct attribute

descriptions (which may be seen as qualitatively described attributes rather than quantitative attributes within BWS and DCE tasks) may add to the difficulty of individuals providing information on the importance of attributes/levels in direct valuation tasks. This suggests that the direct valuation method is more appropriate for areas where respondents will be aware of the value of attributes, and attribute/level descriptions are unambiguous.

These arguments might partly explain why respondents perceived the BWS case I tasks as easier than the DCE binary choice in the GP study. In addition, the nature of the attributes and their description (*Achieved or Not*) in the simple DCE tasks (*Yes or No*) might also help the respondents to easily express the importance they attach to the attributes, thus explaining the convergent validity between the BWS case I and DCE binary choice approaches. For the Glaucoma study, the attributes described negative consequences of the disease and its treatment on the quality of life in a qualitative way. Further, the interaction between the different attributes might have a significant influence on the way the respondents valued the attributes. These potential interactions are not explicitly captured by any attribute and the individuals are not necessarily able to carefully describe their influence on their decisions. Then a 'direct' valuation approach such as BWS seems less appropriate to investigate the importance (values) the respondents attach to the attributes. This might explain why the DCE pairwise choice approach was more able than the and the BWS case II approach to verify the a priori assumptions on the individuals' preferences and to satisfy the theoretical condition of stability.

This argument of unconscious decision making does not mean that the BWS case II approach led to invalid results for the Glaucoma study. Given the concept of importance can be multi-dimensional (Myers & Alpert, 1968), differences between measurement approaches can also be due to differences in the dimensions measured (Van Ittersum et al, 2007). We leave this issue for further research.

The second potential explanation of the similarities/discrepancies between the BWS and DCE approaches would be the underlying decision-making/behavioural models. It is still not clear to what extent DCE tasks based on utility maximisation (i.e. internal evaluation of the utility/desirability of each alternative separately and then comparison of the utility score of each alternative) are in line with the BWS tasks based on the principle of attributes-pair identification (i.e. internal evaluation of the utility of each attribute separately and then selection of the pair of attributes with the largest perceptual difference). Other choice models can also be used to analyse the best and worst decisions. However the empirical cognitive processes underlying these choices remain largely unknown, making it difficult to state under which conditions the BWS and DCE methods are theoretically equivalent. Choosing and rejecting options can lead to different weights (Shafir, 1993). Respondents may give more importance to attributes positively framed when they are asked to accept the options and more importance to negative attributes when they are asked to reject the options. In this perspective, our results showed that the respondents provided significantly more stable worst choices than best ones.

This argument of behavioural compatibility might explain the similarities of the results obtained with the BWS case I and DCE binary choice approaches. In the GP study discrepancies are less likely to occur for two main reasons. The DCE tasks are much closer to

the BWS tasks, since they included only one option and respondents had to make within-option trade-offs instead of between-options trade-offs. The two extreme levels used to describe the attributes in the DCE tasks favour trade-offs based on the sole importance of the attributes and thus provided information similar to the one obtained in the BWS tasks. Given these two reasons, we supposed that respondents were more likely to use the same decision rules to fulfil the DCE and BWS tasks and to process the information on the same way.

This study is not exempt from limitations. First, all the choice tasks were completed by respondents in a given rather than randomized order. In the Glaucoma study, each BWS task appeared immediately after its corresponding DCE task and in the GP study all the BWS tasks were completed before the DCE tasks. Such designs have the potential to introduce order bias in the comparisons, assuming that the respondents tried to be internally coherent in their different choices. This potential order effect would act in favour of greater similarities in the results of the two methods, decreasing our ability to detect differences. However, the comparisons made between the BWS case II and DCE pairwise choice approaches rule out this possibility, since strong discrepancies occurred. But for the comparisons between the BWS case I and DCE binary choice approaches, the order effect cannot be precluded.

Second, the designs used in the case studies might have introduced biases in the comparisons. For the Glaucoma study, each respondent fulfilled 32 DCE tasks and 32 BWS tasks, a potentially large number of tasks for a choice experiment. This relatively high number of tasks might have exaggerated the discrepancies between the two approaches because of fatigue effect and use of simplifying heuristics. We leave this issue for further research. For example, one could analyse how the consistency of the best and worst choices (i.e. variance of errors) evolve along the tasks.

In conclusion, this **s**tudy compared variants of the BWS and DCE approaches in different health contexts. The results showed that the BWS case I approach is easier for the respondents than the DCE binary choice approach. Overall these two approaches are close to each other and provide similar results on the individuals' preferences. This first comparison supports the idea that these two approaches are substitute for each other. The second comparison made between the BWS case II and DCE pairwise choice approaches are in favour of this latter approach, which is more able to satisfy theoretical requirements. Importantly, these two approaches lead to qualitatively different results about individuals' preferences. Then the BWS case II and DCE pairwise choice approaches cannot be seen as substitute for each other. Further research is needed to know if these two approaches provide complementary results or if one of them provides less misleading results.

Bibliography

Al-Janabi, H., T N. Flynn, and J Coast. 2010. "Estimation of a Preference-Based Carer Experience Scale." *Medical Decision Making* 31 (3): 458–468. doi:10.1177/0272989X10381280.

Burr, J M., M Kilonzo, L Vale, and M Ryan. 2007. "Developing a Preference-Based Glaucoma Utility Index Using a Discrete Choice Experiment." *Optometry and Vision Science* 84 (8): 797–808.

Coast, J., C. Salisbury, D. de Berker, A. Noble, S. Horrocks, T.J. Peters, and T N. Flynn. 2006. "Preferences for Aspects of a Dermatology Consultation." *British Journal of Dermatology* 155 (2): 387–392. doi:10.1111/j.1365-2133.2006.07328.x.

De Bekker-Grob, E W., M Ryan, and K Gerard. 2012. "Discrete Choice Experiments in Health Economics: A Review of the Literature." *Health Economics* 21 (2): 145–172. doi:10.1002/hec.1697.

Dijksterhuis, Ap, PK. Smith, RB. van Baaren, and D H.J. Wigboldus. 2005. "The Unconscious Consumer: Effects of Environment on Consumer Behavior." *Journal of Consumer Psychology* 15 (3): 193–202.

Finn, A, and J J. Louviere. 1992. "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety." *Journal of Public Policy & Marketing* 11 (2): 12–25.

Flynn, TN., JJ. Louviere, TJ. Peters, and J Coast. 2007. "Best-Worst Scaling: What It Can Do for Health Care Research and How to Do It." *Journal of Health Economics* 26 (1): 171–189. doi:10.1016/j.jhealeco.2006.04.002.

Flynn, TN., T.J. Peters, and J Coast. 2011. "Quantifying Response Shift or Adaptation Effects in Quality of Life by Synthesising Best-Worst Scaling and Discrete Choice Data". International Choice Modelling Conference, Leeds, UK.

Gul, F, and Wolfgang Pr. 2005. "The Case for Mindless Economics". Princeton University.

Hougaard, Jens Leth, Tue Tjur, and Lars Peter Østerdal. 2011. "On the Meaningfulness of Testing Preference Axioms in Stated Preference Discrete Choice Experiments." *The European Journal of Health Economics* 13 (4): 409–417. doi:10.1007/s10198-011-0312-4.

Houthakker, H.S. 1950. "Revealed Preference and the Utility Function." *Economica* 17 (66): 159–174.

Lagerkvist, CJ. 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.

Lancsar, E, J Louviere, and T Flynn. 2007. "Several Methods to Investigate Relative Attribute Impact in Stated Preference Experiments." *Social Science & Medicine* 64 (8): 1738–1753. doi:10.1016/j.socscimed.2006.12.007.

Louviere, J.J., and G.G. Woodworth. 1990. "Best-Worst Scaling: A Model for Largest Difference Judgments". Faculty of Business, University of Alberta.

Louviere, JJ., and T Islam. 2008. "A Comparison of Importance Weights and Willingness-to-Pay Measures Derived from Choice-Based Conjoint, Constant Sum Scales and Best–worst Scaling." *Journal of Business Research* 61 (9): 903–911. doi:10.1016/j.jbusres.2006.11.010.

McFadden, D. 1999. "Rationality for Economists?" *Journal of Risk and Uncertainty* 19 (1-3): 73–105.

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): 367–382. doi:10.1016/S0167-4870(02)00081-8.

San Miguel, M Ryan, and Me Amaya-Amaya. 2005. "'Irrational' Stated Preferences: A Quantitative and Qualitative Investigation." *Health Economics* 14 (3): 307–322. doi:10.1002/hec.912.

Mueller, S, Larry L, and J. Louviere. 2009. "What You See May Not Be What You Get: Asking Consumers What Matters May Not Reflect What They Choose." *Marketing Letters* 21 (4) (November 20): 335–350. doi:10.1007/s11002-009-9098-x.

Myers, JH., and MI. Alpert. 1968. "Determinant Buying Attitudes: Meaning and Measurement." *Journal of Marketing* 32 (4): 13–20.

Potoglou, D, P Burge, T Flynn, A Netten, J Malley, J Forder, and J. Brazier. 2011. "Best–worst Scaling vs. Discrete Choice Experiments: An Empirical Comparison Using Social Care Data." *Social Science & Medicine* 72 (10): 1717–1727. doi:10.1016/j.socscimed.2011.03.027.

Ryan, M, and K 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, M, K Gerard, and M Amaya-Amaya. 2008. *Using Discrete Choice Experiments to Value Health and Health Care*. Dordrecht: Springer.

Ryan, M, V Watson, and V Entwistle. 2009. "Rationalising the 'irrational': A Think Aloud Study of Discrete Choice Experiment Responses." *Health Economics* 18 (3): 321–336. doi:10.1002/hec.1369.

Samuelson, P.A. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5 (17): 61–71.

Shafir, Eldar, Itamar Simonson, and Amos Tversky. 1993. "Reason-Based Choice." *Cognition* 49: 11–36.

Siegel, Sydney, and N. John Castellan Jr. 1988. *Nonparametric Statistics for The Behavioral Sciences*. 2nd ed. McGraw-Hill.

Van Ittersum, Koert, Joost M.E. Pennings, Brian Wansink, and Hans C.M. van Trijp. 2007. "The Validity of Attribute-Importance Measurement: A Review." *Journal of Business Research* 60 (11): 1177–1190. doi:10.1016/j.jbusres.2007.04.001.

Wagner, EH., BT. Austin, C Davis, M Hindmarsh, J Schaefer, and A Bonomi. 2001. "Improving Chronic Illness Care: Translating Evidence into Action." *Health Affairs* 20 (6): 64–78.

Whitty, J, J Ratcliffe, Gang Chen, and P Scuffham. 2013. "A Comparison of Discrete Choice and Best Worst Scaling Methods to Assess Australian Public Preferences for the Funding of New Health Technologies". International Choice Modelling Conference, Sydney, Australia.

Witt, J, A Scott, and R Osborne. 2009. "Designing Choice Experiments with Many Attributes. An Application to Setting Priorities for Orthopaedic Waiting Lists." *Health Economics* 18 (6): 681–696. doi:10.1002/hec.1396.

Appendix 1. Review of the BWS studies published in health

Article	Objective	Domain	Sample	Nbr respondents	BWS method	Objects	Nbr tasks	Task format	Scale
Gallego et al (2012)	To measure impact of emerging technologies on health outcomes	Hepatocellular carcinoma	Health professionals	120	Case I	11 items	11	5 items	Impact
Wang et al (2011)	To identify the importance of factors affecting the medical students' choices of residency programs	Professional choices	Medical students	339	Case I	13 items	13	4 items	Importance
Mazanov et al (2012)	To determine the importance of health values in sport	Health principles	Population	168	Case I	11 items	11	5 items	Importance
Marti (2012)	To assess the adolescents' concerns for adverse effects of tobacco use	Tobacco use	Population	376	Case I	15 items	16	5;7;9;11 items	Deterrence
Louviere & Flynn (2010)	To evaluate the importance of Australian healthcare reform principles	Healthcare reform	Population	204	Case I	15 items	15	7;8 items	Importance
Al-Janabi et al (2011)	To construct a quality of life measure specific to carer	Quality of life	Informal carer	397	Case II	6 attributes (6x3L)	18	6 items	Desirability
Flynn et al (2008)	To compare different approaches to analyze the BWS data	Dermatology consultation	Patients	55	Case II	4 attributes (1x4L and 3x2L)	16	4 items	Desirability
Molassiotis et al (2012)	To measure utilities associated with delivery of a symptom management intervention	Lung cancer	Patients	87	Case II	4 attributes (4x2L)	16	4 items	Desirability
Ratcliffe et al (2012)	To obtain adolescent-specific values for the Child Health Utility 9D	Health Utility	Population	590	Case II	9 attributes (9x5L)	10	9 items	Desirability
Günther et al (2010)	To quantify the preferences of young physicians for practice establishment	Professional choices	Health professionals	5026	Case II	6 attributes (4x3L and 2x4L)	24	6 items	Desirability
Potoglou et al (2011)	To compare the DCE and BWS methods	Social care	Population	300	Case II	9 atributes (8x4L and 1x2L)	12	9 items	Desirability
van Dijk et al (2013)	To compare the DCE and BWS methods	Hip arthroplasty	Population	429	Case II	5 attributes (5x4L)	8	5 items	Desirability
Severin et al (2013)	To measure preferences for prioritising alternative clinical interventions	Genetic test	Health professionals	26	Case II	6 attributes	?	6 items	Desirability
Coast et al (2006)	To quantify preferences for different aspects of care	Dermatology service	Population	99	Case II	4 attributes (1x4L and 3x2L)	16	4 items	Desirability
Whitty et al (2013)	To compare the validity and acceptability of DCE and BWS methods	Health Technology Assessment	Population	900	Case II	7 attributes (3x2L, 1x3L, 2x4L, 1x6L)	6	7 items	Importance
Lancsar et al (2012)	To illustrate the BWS case 3 approach in health	Cardiac arrest	Population	898	Case III	3 attributes (2x4L and 1x2L)	16	5 options	Desirability
Xie et al (2013)	To compare the feasibility, reliability and results of DCE and BWS methods	EQ-5D-5L	Population	177	Case III	5 attributes (5x5L)	16	3 options	Desirability