VALUE IN HEALTH ▮ (2016) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

An Empirical Comparison of Discrete Choice Experiment and Best-Worst Scaling to Estimate Stakeholders' Risk Tolerance for Hip Replacement Surgery

Joris D. van Dijk, MSc[1], *Catharina G.M. Groothuis-Oudshoorn, PhD*[1,*], *Deborah A. Marshall, PhD*[2], *Maarten J. IJzerman, PhD*[1]

[1]*Department of Health Technology & Services Research, MIRA Institute for Biomedical Technology & Technical Medicine, University of Twente, Enschede, The Netherlands;* [2]*Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada*

ABSTRACT

Background: Previous studies have been inconclusive regarding the validity and reliability of preference elicitation methods. **Objective:** The aim of this study was to compare the metrics obtained from a discrete choice experiment (DCE) and profile-case best-worst scaling (BWS) with respect to hip replacement. **Methods:** We surveyed the general US population of men aged 45 to 65 years, and potentially eligible for hip replacement surgery. The survey included sociodemographic questions, eight DCE questions, and twelve BWS questions. Attributes were the probability of a first and second revision, pain relief, ability to participate in sports and perform daily activities, and length of hospital stay. Conditional logit analysis was used to estimate attribute weights, level preferences, and the maximum acceptable risk (MAR) for undergoing revision surgery in six hypothetical treatment scenarios with different attribute levels. **Results:** A total of 429 (96%) respondents were included. Comparable attribute weights and level preferences were found for both BWS and DCE. Preferences were greatest for hip replacement surgery with high pain relief and the ability to participate in sports and perform daily activities. Although the estimated MARs for revision surgery followed the same trend, the MARs were systematically higher in five of the six scenarios using DCE. **Conclusions:** This study confirms previous findings that BWS or DCEs are comparable in estimating attribute weights and level preferences. However, the risk tolerance threshold based on the estimation of MAR differs between these methods, possibly leading to inconsistency in comparing treatment scenarios.

Keywords: benefit-risk assessment, discrete choice experiment, best-worst scaling, patient preference, preference elicitation.

Introduction

Over the last few years, regulatory agencies increasingly consider stated-preference methods to allow a more explicit and transparent judgment based on preferences of patients and other stakeholders [1,2]. The implicit assumption is that considering patient preferences will increase the quality of regulatory decisions and communication to a large group of stakeholders [3,4]. Among others, Johnson et al. [5] and Phillips et al. [6] suggested that stated-preference surveys are one of the most reliable and valid techniques available for quantifying patient preferences [5–8]. Yet, the use of patient-preference data raises several concerns for policymakers. In particular, the validity and reliability of the available preference elicitation methods and the consistency and comparability of the results produced [9].

Several methods can be used to elicit patient preferences including matching methods, conjoint analysis, and multicriteria decision analysis. Weernink et al. [10] provided a comprehensive review of preference elicitation methods for regulatory decision making and concluded that matching methods and conjoint analysis fulfill the requirements to support regulatory decision making.

The most commonly used stated-preference format in health care is the discrete choice experiment (DCE) [11,12]. In a DCE, respondents are presented with several choice sets. Each choice set consists of multiple hypothetical profiles consisting of a fixed set of criteria (attributes), with varying values (levels) between the profiles. See also Ryan et al. [13] for more details.

Best-worse scaling (BWS) has been introduced as an alternative to DCEs and creates the ability to compare the relative impact of attributes [14,15]. There are three types of BWS: BWS case 1, 2, and 3, in which the difference is mainly determined by the use of attributes or attributes levels and the use of single or multiple profiles. The BWS case 2 is most comparable to the dual-choice DCE [15,16]. In BWS case 2 or profile-case BWS, respondents are presented with several choice sets consisting of one

* *Address correspondence to:* Catharina G.M. Groothuis-Oudshoorn, Department of Health Technology & Services Research, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

E-mail: c.g.m.oudshoorn@utwente.nl.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2015.12.020

profile only presenting attribute levels comparable to a DCE. Respondents are then asked to select the most (best) and least (worst) preferred criterion. See Flynn [17] for more details.

Previous work in reviewing the literature and work performed by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task forces has proposed best practices and guidelines for conducting stated-preference studies, in particular for DCE [7,13,17–19]. However, although the use of BWS is growing, there is less experience with BWS than with DCE. To date, only limited empiric research has been conducted to determine the validity and reliability of both methods with respect to estimating preferences and the studies are inconclusive yet. Potoglou et al. [20] and Severin et al. [21] both conclude that BWS and DCE weights for the attributes differed in detail, yet the evaluation patterns are consistent. Whitty et al. [22] concluded that DCE formats are valid and reliable, where the validity of BWS formats is less definitive. In a study by Xie et al. [23], preferences for the five-level EuroQol five-dimensional questionnaire were obtained using DCE and multiprofile BWS. They concluded that the DCE was more valid and reliable, yet variance of the preferences with DCE was larger [23]. Whitty et al. [24] observed differences in the decision-making process of respondents in a head-to-head comparison of BWS and DCE and expressed concerns about uncertainty regarding the use of these methods for priority setting.

Hence, the present study was designed to specifically compare profile-case BWS with DCE and to evaluate whether the metrics estimated using these methods could lead to different regulatory decisions. Such decision sensitivity expresses whether the decision made depends on the method used and is a relevant metric for policymakers [25].

The present study explicitly considers the choice for type of hip replacement procedure, where hip resurfacing arthroplasty (HRA) and total hip arthroplasty (THA) are the options. These treatments were chosen as a case because of the clinical equipoise in certain patient groups [26,27]. That is, HRA is considered an alternative to THA, where HRA is associated with additional clinical benefits at the expense of the higher risk of revision procedures [26,28–30].

In addition to preference weights for attributes, another metric that 1) expresses the risk tolerance for a hip replacement procedure; and 2) may inform a decision maker is the maximum acceptable risk (MAR). The MAR was introduced by Johnson et al. [5], and it expresses the risk that patients are willing to take to gain a certain amount of benefit [31]. The MAR is theoretically higher if the benefits produced by the intervention increase. Hence, the aim of this study was to compare the metrics obtained from DCE and BWS and to determine whether these methods could lead to different regulatory decisions, based on the MAR, with respect to hip replacement.

Methods

Study Population

A study sample was selected in June 2012 from an international Internet panel maintained by Survey Sampling International of the general population in the United States. Only men aged between 45 and 65 years were eligible to participate because of the clinical equipoise in these patient groups. A minimum sample size of at least 300 was required, based on the recommendations for quantitative research by Marshall et al. [11], Orme [32], and Johnson et al. [33].

Attribute Selection

Potentially relevant attributes related to THA and HRA were identified on the basis of multiple National Joint Replacement Registries and systematic reviews [26,28,29,34–39]. All attributes were judged by two medical researchers for their impact on a possible decision in choosing between THA and HRA. The selection of the attributes was based on the occurrence in literature, possible effects on daily activities, relevance for decision making for patients, and the extent to which they were distinctive for the choice between THA and HRA. Five attributes were selected in the final set: 1) the probability of a first revision in 7 years; 2) the probability of a failing second implant (after replacement surgery) per 7 years; 3) pain relief; 4) the ability to perform moderate or heavy daily tasks and participate in sports; and 5) hospital stay. The first two attributes were framed as risk, whereas the latter three were framed as benefits.

Four attribute levels were chosen for all five attributes on the basis of literature reviews and registries [26,28,29,34–39]. Revision rates were based on the reported outcomes for men between 45 and 65 years in the Australian annual National Joint Replacement Registry and the National Joint Replacement Registry of England and Wales [26,28]. These registries are members of the International Consortium of Registries and are the largest registries available with the longest follow-up times. The levels for each of the benefits were based on systematic reviews [35–39] and the Alberta Hip Improvement project [29,34], as presented in Table 1.

Overall Survey Design

The survey consisted of three parts: an introduction explaining the relevance of the survey followed by several sociodemographic questions and 8 DCE and 12 BWS choice tasks. The order of both types of choice sets was randomized between respondents to prevent ordering effects. Respondents were asked their preference for multiple hypothetical hip replacement therapies, all based on HRA and THA. A short introduction, including an example task, was given before each type of choice set (as illustrated in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2015.12.020). Respondents rated the difficulty of the choice tasks associated with BWS and DCE with a number ranging from 1 (very easy) to 5 (very difficult). Respondents had unlimited time to complete the survey. The time spent on the 12 BWS tasks and on the eight DCE tasks, excluding reading of the introductions, was recorded for each respondent.

Design of the DCE Choice Sets

A full-profile DCE with fractional factorial design using triplets with a balanced overlap strategy and orthogonal design is used [13,33,40,41]. A random sample of 200 versions of the questionnaire was constructed consisting of a random selection of eight choice tasks out of all possible profiles [15]. The number of tasks was based on the minimal required number of choice tasks given the amount of attributes and levels in this study. The choice sets were obtained randomly with Sawtooth's (Sequim, WA) design algorithm using SSI web 8.0.2 and designed for main effects only. The three alternatives in each choice task were unlabelled, and the attribute order was kept constant for each respondent in all the choice sets but was randomized between respondents.

Design of the BWS Choice Sets

Full-profile BWS choice tasks were created using an orthogonal design to ensure level balance (as illustrated in Supplemental Materials). For the survey, a random sample of 200 versions of the BWS choice sets was constructed consisting of 12 choice tasks out of the 1024 possible profiles [15]. This is considered the minimal required number of tasks given the amount of attributes and levels in this study [18]. The profiles were obtained by varying all attribute levels using the algorithm of Sawtooth with 2000 iterations to gain level balance. The position of the attributes in

Table 1 – Attributes and levels for eliciting preferences, based on literature review.

Attributes	Levels (1–4)		Performance in literature HRA and THA
Number of people needing an implant replacement surgery in 7 y (revision)	1	0 of 100 (0%)	Based on performance in the National Joint Replacement Registries of Australia and England and Wales: varying from 2.8% to 4.8% revision for THA in men aged between 55 and 75 y and from 4.1% to 7.1% for HRA [26,28]
	2	3 of 100 (3%)	
	3	6 of 100 (6%)	
	4	9 of 100 (9%)	
Number of people with a failing second implant (after replacement surgery) in 7 y	1	0 of 100 (0%)	Based on performance reported in the Australian annual National Joint Replacement Registries; ±31% in HRA and 18% in THA per 7 y [28], although some literature reports no significant differences [37,38]
	2	10 of 100 (10%)	
	3	20 of 100 (20%)	
	4	30 of 100 (30%)	
The moderate pain is decreased after treatment to:	1	No pain	No significant differences reported between HRA and THA and for both an increase in Harris hip scores corresponding with an improvement to no pain is reported in the Alberta Joint report [34–36]
	2	Slight pain	
	3	Mild pain	
	4	Moderate pain	
Number of people able to perform moderate or heavy tasks and participate in sports after treatment	1	10 of 10 (100%)	Based on results reported by Vendittoli et al. [39]: 72% for patients with HRA and 39% for patients with THA
	2	7 of 10 (70%)	
	3	4 of 10 (40%)	
	4	0 of 10 (0%)	
Average hospital stay is	1	3 d	Longer stay for THA although different outcomes reported; 4.1 d for HRA and 4.7 d for THA and 1.4 d less for HRA [34,35]
	2	4 d	
	3	5 d	
	4	6 d	

HRA, hip resurfacing arthroplasty; THA, total hip arthroplasty.

the BWS choice tasks varied randomly so respondents had to reconsider them carefully for each new choice task.

Data and Statistical Analysis

Respondents who were considered nontraders were excluded from the analysis. *Nontraders* were defined as respondents who had chosen the same profile position in 7 of 8 choice tasks for the DCE, and/or if they had selected the same attribute position (either best or worst) in more than 10 of the 12 tasks in the BWS choice set. This criterion is justified by the complete randomization of the choice tasks and profile formats.

We assume that the utility of an alternative can be modeled as a linear function of the attributes and levels, namely,

$$U = \sum_{i=1}^{20} \beta_i x_i + \varepsilon,$$

where β_i are the part-worth utility parameters for all levels of the attributes, x_i is 1 if the associated level of a certain attribute is present in the particular alternative, and ε is the random error, representing the individual variation in preferences.

A conditional logit model was used to estimate the part-worth utilities for both the DCE and the BWS data as described by Flynn et al. [15]. Dummy coding was used for both methods. For the DCE data, the attribute level with the lowest preference was taken as the reference level for each attribute (level 4). For BWS, the attribute level with the lowest preference (level 4) of the least important attribute was taken as the reference level. The relative change in part-worth utilities for changing levels and attribute importance were compared between both methods. The confidence intervals of MARs were calculated using a bootstrap method (50 replications).

Estimation of the Mar

The risk tolerance, that is, the revision risk people accepted for an incremental benefit, was quantified by MAR [31]. MAR represents the amount of risk associated with a first revision for which HRA equals the part-worth utility or preference for THA. The attribute levels associated with THA and the levels of the six scenarios of HRA, as presented in Table 2, were based on reported and hypothetical performances [26,28,34,39].

The expected utilities were estimated by calculating the difference in all (five for THA and four for HRA) part-worth utilities associated with the levels in the scenarios, as described by Train [42] and Louviere et al. [43].

For example, the total part-worth utilities of THA (U_{THA}) and HRA (U_{HRA}) were determined by adding the part-worth utilities of every attribute β_{att} and the corresponding attribute-level performance (see Table 2).

$$U_{THA} = \beta_{1,L} + \beta_{2,L} + \beta_{3,L} + \beta_{4,L} + \beta_{5,L} = \beta_{1,4.1\%\ \text{risk first revision}} + \beta_{2,30\%\ \text{risk second revision}} + \beta_{3,\text{slightpain}} + \beta_{4,40\%\ \text{to perform tasks and participate in sports}} + \beta_{5,4.7\ \text{days}}$$

$$U_{HRA} = \beta_{2,L} + \beta_{3,L} + \beta_{4,L} + \beta_{5,L}$$

where $\beta_{2,30\%}$ is described by the part-worth utility corresponding to 30% risk on second revision. The difference $U_{THA} - U_{HRA}$ is the part-worth utility associated with MAR for a first revision, as presented in Table 3. Consecutively, this part-worth utility can be converted to the associated risk percentage using linear interpolation.

Statistical Analysis

Patient-specific parameters and characteristics were determined as mean ± SD using Stata (StataSE, version 12.0). The time to complete the choice sets of both methods and the reported difficulties were compared using the Wilcoxon signed rank test. The influence of education level on the time to complete a choice set and the reported difficulty to complete the choice sets were tested using the Kruskal Wallis test.

The level of statistical significance was set to 0.05 for all statistical analyses.

Results

Of all 541 respondents who opened the online survey, 447 completed all questions. Finally, 429 respondents made consistent choices and were included in the analysis. The baseline

Table 2 – Maximum acceptable risk (MAR) of a first revision of HRA using DCE and BWS, including six scenarios with varying outcomes for HRA.

Attributes	Scenario						
	Base case (THA)	1	2	3	4	5	6
Probability of a first revision in 7 y (%)	4.1 [26]	MAR*	MAR*	MAR*	MAR*	MAR*	MAR*
Probability of a second revision in 7 y (%)	±30 [28]	10	30	30	30	20	20
Pain after the procedure	Slight [34–36]	Slight	*None*	Slight	Slight	Slight	*None*
Ability to perform daily tasks (%)	±40 [39]	40	40	*70*	40	*70*	*70*
Average hospital stay (d)	4.7 [34]	4.7	4.7	4.7	4	4.7	4.7
MAR* chance first revision in 7 y DCE (%)	–	41.1 ± 33.0	28.1 ± 20.6	25.1 ± 17.8	3.0 ± 0.2	34.1 ± 26.6	19.0 ± 3.8
MAR* chance first revision in 7 y BWS (%)	–	23.6 ± 8.1	20.0 ± 6.6	16.3 ± 4.8	5.3 ± 0.5	8.4 ± 1.7	5.6 ± 0.6

Note. MARs for a first revision were calculated for six scenarios with varying attribute levels. The attribute levels that are differing between the base-case scenario (THA) and the HRA scenarios are given in italic.

BWS, best-worst scaling; DCE, discrete choice experiment; HRA, hip resurfacing arthroplasty; THA, total hip arthroplasty.

* Estimated MAR of a first revision for each scenario in which HRA is equally preferred as THA.

characteristics and history of all included respondents are summarized in Table 4.

Using weights estimated with DCE, it was found that respondents preferred a hip replacement surgery resulting in the highest probability to perform daily tasks or participate in sports, pain relief, the lowest chance on a first and a second revision, and a minimal hospital stay (Table 3).

The ordering of each of the levels was similar for both methods, except for the third level of hospital stay as estimated with the DCE. Although the rank-order was similar, the relative

Table 3 – Attribute levels and respondent weights for both the DCE and the BWS.

Attribute levels	DCE ($R^2 = 0.22$)		BWS ($R^2 = 0.20$)	
	β coefficient	95% CI	β coefficient	95% CI
Chance first revision in 7 y (%)				
0	0.54*	0.42 to 0.66	1.86*	1.72–2.01
3	0.31*	0.19 to 0.43	0.42*	0.28–0.56
6	0.07†	0.06 to 0.19	0.02†	0.12 to 0.16
9	Reference		0.14†	0.27 to 0.00
Chance second revision in 7 y (%)				
0	0.93*	0.81 to 1.05	1.52*	1.38 to 1.66
10	0.56*	0.43 to 0.68	−0.27*	−0.40 to −0.13
20	0.20*	0.07 to 0.33	−0.45*	−0.59 to −0.32
30	Reference		−0.65*	−0.78 to −0.51
Number of people able to perform moderate or heavy tasks and participate in sports after treatment (%)				
100	1.68*	1.55 to 1.82	2.11*	1.96–2.25
70	1.15*	1.02 to 1.28	1.11*	0.97–1.25
40	0.58*	0.44 to 0.71	0.96*	0.82–1.10
0	Reference		−0.60*	−0.74 to −0.46
Moderate pain decrease to				
None	1.50*	1.37 to 1.63	2.44*	2.29–2.59
Slight	0.86*	0.73 to 0.99	1.81*	1.67–1.95
Mild	0.64*	0.50 to 0.77	0.92*	0.78–1.05
Moderate	Reference		−0.37*	−0.51 to −0.24
Average hospital stay (d)				
3	0.14‡	0.02 to 0.26	0.62*	0.48–0.77
4	0.12†	−0.01 to 0.24	0.49*	0.34–0.63
5	−0.02†	−0.15 to 0.10	0.26*	0.11–0.40
6	Reference		Reference	

CI, confidence interval.

* $P < 0.001$.

† $P > 0.05$.

‡ $P \leq 0.05$.

Table 4 – Self-reported baseline characteristics and history of all included respondents.

Characteristic	n	%
Respondents opened survey	541	126
Respondents completed all questions	447	104
Consistent DCE and BWS	429	100
Consistent DCE	432	
Consistent BWS	444	
Age (y)		
45–55	213	50
56–65	216	50
Health status		
No problems	137	32
Slight problems	167	39
Moderate problems	99	23
Severe problems	26	6
Familiar with hip complaints		
Yes, self	147	34
Yes, via other	72	17
No	210	49
Education level		
High school, public or primary school	115	27
Trade or technical qualification	82	19
College or university	232	54
Marital status		
Not single	291	68
Single	138	32

BWS, best-worst scaling; DCE, discrete choice experiment.

change in utility for attaining a higher attribute level differed between both methods for all attributes, as illustrated in Figure 1. BWS was found to be more sensitive to a change in risks in the lowest levels (from level 1 to 2). For instance, the estimated relative utility change (compared with the disutility of shifting from the lowest to the highest level on performing daily tasks or participating in sports) for shifting from the first (0%) to the second level (3%) for the risk on a first revision was 38% higher in case of BWS than of DCE. Yet, shifting from the second (3%) to the fourth level (9%) was comparable between BWS (20%) and DCE (19%) for this attribute. Moreover, comparable utility changes between BWS and DCE were observed for the attributes pain relief and the ability to perform moderate tasks and participate in sports (Figure 2).

Differences were also observed between the calculated MARs for both methods, as presented in Table 2. The use of BWS resulted in lower MARs for a first revision surgery in five of the six scenarios, ranging from 6% to 24% versus 19% to 41% when using DCE. Only when the length of hospital stay was changed (scenario 4), the BWS predictions of MAR were higher (5%) compared with DCE (3%). However, because of the small disutility difference for higher risk levels, these estimates of MARs have large standard errors.

The median time respondents needed to complete the choice sets was shorter for DCE (192 s) than for BWS (335 s) ($P < 0.001$). No significant differences in time to complete the choice sets between the education groups were found for the DCE choice sets (198, 193, and 189 s for the highest, middle, and lowest education groups, respectively; $P = 0.50$). Yet, differences were observed between the education groups when completing the BWS choice sets (350, 389, and 320 s, respectively; $P = 0.02$).

The mean completion difficulty of the DCE was lower (2.5; 95% confidence interval 2.4–2.6) than for BWS (2.6; 95% confidence interval 2.5–2.7) but not significant ($P = 0.06$). Also, no significant differences were found between the education groups rating the choice set difficulty ($P = 0.21$).

Discussion

This study investigated the difference between profile-case BWS and DCE for estimating the attribute weights and MARs for different hip replacement surgery scenarios. We conclude that the pattern of attribute weights and levels is similar for both DCE and BWS. MARs estimated for a first revision for six scenarios using these preference weights followed a similar pattern estimated with BWS and DCE data. Yet MARs estimated with DCE are systematically higher. For extreme scenarios, MARs can be estimated with DCE and BWS quite consistently. However, when differences between the base-case scenario and the postulated scenario are smaller, that is, when considering scenario 4 (see Table 2), the different methods produce different results. Because these scenarios are more likely to occur, the use of either DCE or BWS will produce different results when it comes to regulatory decisions.

Although we are the first to compare the MAR obtained with either BWS or DCE, similar patterns in attribute weights and level preferences when using either BWS or DCE are in agreement with previous studies. Potoglou et al. [20] were the first to compare BWS and DCE and also showed similar estimated attribute-level weights and attribute importance using DCE and BWS. This also is in agreement with a study of Severin et al. [21] who also showed comparable preference patterns. A recent study by Whitty et al. [22] comparing BWS and DCE in cases of priority setting also observed similar trends regarding attribute weights, but encountered more uncertainty in the BWS choice sets . Another study by Whitty et al. [24] showed differences in the relative preference weights and therefore raised concerns for using one or both methods in a priority-setting context. These differences in preference weights between BWS and DCE might be due to the health decision context. BWS might provide similar findings as DCE only when assessing personal health preference instead of public preferences, as suggested by Whitty et al. [24].

Previous studies have suggested that BWS is less cognitively demanding than DCE [17,20]. Although the present study was not designed to investigate the cognitive burden of either method,

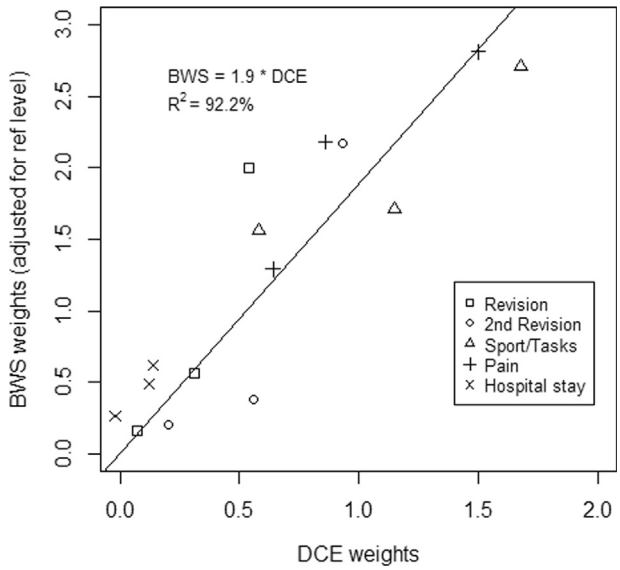

Fig. 1 – Comparison of estimated preference weights by using discrete choice experiment (DCE) or profile-case best-worst scaling (BWS). The solid line represents the fit, and both the fitted result and the coefficient of determination are shown.

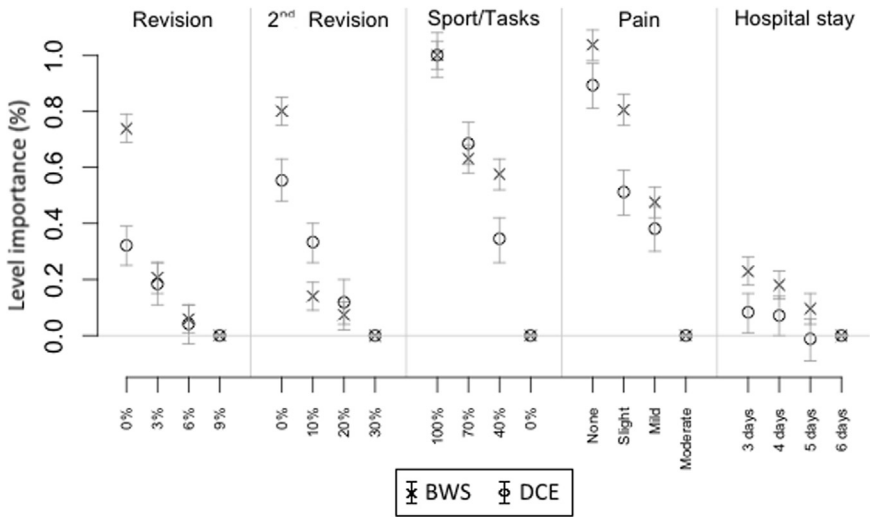

Fig. 2 – Estimated part-worth utilities of best-worst scaling (BWS) and discrete choice experiment (DCE) for the four levels of all five attributes rescaled to the largest disutility (ability to participate in sports and perform daily tasks). The mean estimates are presented including their 95% confidence intervals.

respondents rated the BWS choice sets as more difficult. It also took more time to complete the BWS choice sets, but this figure needs to be balanced against the higher number of questions (12 in the BWS compared with 8 in the DCE). The findings of the cognitive burden are in agreement with the studies of Whitty et al. [22,24] and Severin et al. [21], who reported less difficulty completing DCE than BWS choice tasks. It is hypothesized that the greater difficulty could be due to the more unrealistic choices and fewer possibilities for simplification strategies. Another argument may be the changing order of attributes in each BWS task, compared with DCE in which this was kept constant.

Although we followed international guidelines and recommendations in the construction of the experimental design of both BWS and DCE, some specific choices and assumptions were made, which could have influenced the generalizability and outcomes of this study [7,13,17,18]. First, the selection of the attributes and their levels is known to influence the estimated preferences for both methods [44]. Although experts were involved in the final selection of the attributes and their levels, it remains arbitrary. Moreover, to limit the number of attributes and levels, hypothetical attribute levels were included to cover the entire range of possible HRA performance outcomes in the future. Yet, because the objective of this study was to compare the metrics obtained from BWS and DCE, it was not considered to influence the study outcomes. Second, the outcomes of both methods might change when assessing a different type of treatment or technique. Although it seems that similar estimated level preferences and attribute importance were found for other applications when using DCE and BWS [20–22], our study also demonstrates that it would be possible to encounter differences between both methods when solely looking at a risk threshold such as a MAR. The underlying reason for this might be that the MAR is determined as a ratio and is based on incremental differences instead of absolute numbers. It therefore holds the same limitations as with, for example, the use of incremental cost-effectiveness ratios [45]. The small differences in the estimated preferences or utilities for the attribute levels including their uncertainties are then combined. This resulted in even larger differences in the estimated MARs with high SDs for all scenarios.

This study illustrates that the use of either DCE or profile-case BWS does not seem to influence the estimated preferences or attribute weights in a personal health context. Both methods to estimate attribute weights or level preferences provide comparable results and can be useful for a qualitative consideration of patient preferences in health care [1]. Yet, caution should be exercised when using these preferences to estimate a ratio or absolute cutoff value, that is, MAR, for decision making. Although it would be easier to have one cutoff value, our results demonstrate that the choice of methodology could then influence the decisions and the results can thus be sensitive to regulatory decisions. Therefore, it is recommended not to consider only the threshold or absolute value but also to incorporate the underlying respondents' preferences in decision making.

Conclusions

This study supports the use of either BWS or DCE in estimating attribute weights and level preferences in a personal health context. Although the calculated MARs on a first revision followed the same trend for both methods, the risk tolerance threshold may vary, possibly leading to inconsistency in discriminating between treatments. Hence, we advise using either BWS or DCE as a decision support tool to better estimate attribute weights and level preferences but caution should be exercised when converting the preferences in a threshold for medical decision making.

Source of financial support: Dr. Marshall is supported by a Canada Research Chair in Health Services and Systems Research and the Arthur J.E. Child Chair Rheumatology Outcomes Research.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2015.12.020 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Report of the CHMP Working Group on Benefit-Risk Assessment Models and Methods. London: European Medicines Agency, CHMP, 2007.

[2] Hauber AB, Fairchild AO, Johnson FR. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. Appl Heal Econ Heal Policy 2013;11:319–29.
[3] Road Map to 2015: The European Medicines Agency's Contribution to Science, Medicines and Health. London: European Medicines Agency, 2010.
[4] Danner M, Hummel M, Volz F, et al. Integrating patients' views into health technology assessment: analytic hierarchy process (AHP) as a method to elicit patient preferences. Int J Technol Assess Health Care 2011;27:369–75.
[5] Johnson FR, Hauber AB, Poulos CM. A Brief Introduction to the Use of Stated-Choice Methods to Measure Preferences for Treatment Benefits and Risks (Publication No. RR-0009-0909). Res Triangle Park, NC: RTI Press, 2009.
[6] Phillips KA, Johnson FR, Maddala T. Measuring what people value: a comparison of "attitude" and "preference" surveys. Health Serv Res 2002;37:1659–79.
[7] Bridges JFP, Hauber AB, Marshall D, et al. Conjoint analysis applications in health–a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health 2011;14:403–13.
[8] Lagarde M, Blaauw D. A review of the application and contribution of discrete choice experiments to inform human resources policy interventions. Hum Resour Health 2009;7:62.
[9] Van Til JA, Ijzerman MJ. Why should regulators consider using patient preferences in benefit-risk assessment? Pharmacoeconomics 2014;32:1–4.
[10] Weernink MGM, Janus SIM, van Til JA, et al. A systematic review to identify the use of preference elicitation methods in healthcare decision making. Pharmaceut Med 2014;28:175–85.
[11] Marshall D, Bridges JFP, Hauber AB, et al. Conjoint analysis applications in health – how are studies being designed and reported? An update on current practice in the published literature between 2005 and 2008. Patient 2010;3:249–56.
[12] de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012;172:145–72.
[13] Ryan M, Amaya-Amaya M, Gerard K. Using Discrete Choice Experiments to Value Health and Health Care. Dordrecht, the Netherlands: Elsevier, 2008.
[14] Finn A, Louviere JJ. Determining the appropriate response to evidence of public concern: the case of food safety. J Public Policy Mark 1992;11:22–5.
[15] Flynn TN, Louviere JJ, Peters TJ, Coast J. Best–worst scaling: what it can do for health care research and how to do it. J Health Econ 2007;26:171–89.
[16] Ratcliffe J, Flynn TN, Terlich F, et al. A pilot study to apply best worst scaling discrete choice experiment methods to obtain adolescent specific values for the Child Health Utility 9D. Flinders Centre for Clinical Change Working Paper 2011/1, Flinders University. South Australia, 2011.
[17] Flynn TN. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Rev Pharmacoecon Outcomes Res 2010;10:259–67.
[18] Orme BK. Accuracy of HB Estimation in MaxDiff Experiments (Sawtooth Software Research Paper Series). Sequim, WA: Sawtooth Software, 2005.
[19] Lancsar E, Louviere J, Flynn TN. Several methods to investigate relative attribute impact in stated preference experiments. Soc Sci Med 2007;64:1738–53.
[20] Potoglou D, Burge P, Flynn TN, et al. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. Soc Sci Med 2011;72:1717–27.
[21] Severin F, Schmidtke J, Mühlbacher A, Rogowski WH. Eliciting preferences for priority setting in genetic testing: a pilot study comparing best-worst scaling and discrete-choice experiments. Eur J Hum Genet 2013;21:1202–8.
[22] Whitty JA, Walker R, Golenko X, Ratcliffe J. A think aloud study comparing the validity and acceptability of discrete choice and best worst scaling methods. PLoS One 2014;9:e90635.
[23] Xie F, Pullenayegum E, Gaebel K, et al. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best-worst scaling? Eur J Health Econ 2014;15:281–8.
[24] Whitty JA, Ratcliffe J, Chen G, Scuffham PA. Australian public preferences for the funding of new health technologies: a comparison of discrete choice and profile case best-worst scaling methods. Med Decis Making 2014;34:638–54.
[25] Felli JC, Hazen GB. Sensitivity analysis and the expected value of perfect information. Med Decis Mak 1998;18:95–109.
[26] National Joint Registry for England and Wales. Annual Report. National Joint Registry, Hertfordshire, England, 2011.
[27] Bozic KJ, Browne J, Dangles CJ, Manner PA, Yates AJ Jr, Weber KL, Boyer KM, Zemaitis P, Woznica A, Turkelson CM, Wies JL. Modern metal-on-metal hip implants. Journal of the American Academy of Orthopaedic Surgeons 2012;20(6):402–6.
[28] Australian Orthopaedic Association National Joint Replacement Registry. Annual Report 2010. Adelaide, Australia: Australian Orthopaedic Association, 2010.
[29] Alberta Bone and Joint Health Institute. Alberta Hip Improvement Project: Preliminary Report on Early Results of Metal-on-Metal Resurfacing for Treatment of Degenerative Hip Disease in Alberta. Alberta Bone and Joint Health Institute, Calgary, Alberta, Canada, 2010.
[30] Garellick G, Kärrholm J, Herberts P. Swedish Hip Arthroplasty Register Annual Report 2010. Gothenburg, Sweden: Swedish Hip Arthroplasty Register, 2011.
[31] Johnson FR. Quantifying Patient Benefit-Risk Tradeoff Preferences. A Brief Introduction. RTI Health Solutions, Durham, NC, 2008:8.
[32] Orme BK. Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research. Madison, WI: Research Publishers LLC, 2009.
[33] Johnson FR, Lancsar E, Marshall D, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value Health 2013;16:3–13.
[34] Alberta Bone and Joint Health Institute. Metal-On-Metal Hip Resurfacing for Young Active Adults with Degenerative Hip Disease. Alberta Bone and Joint Health Institute, Calgary, Alberta, Canada, 2006.
[35] Smith TO, Nichols R, Donell ST, Hing CB. The clinical and radiological outcomes of hip resurfacing versus total hip arthroplasty: a meta-analysis and systematic review. Acta Orthop 2010;81:684–95.
[36] Jiang Y, Zhang K, Die J, et al. A systematic review of modern metal-on-metal total hip resurfacing vs standard total hip arthroplasty in active young patients. J Arthroplasty 2011;26:419–26.
[37] Corten K, Ganz R, Simon JP, Leunig M. Hip resurfacing arthroplasty: current status and future perspectives. Eur Cell Mater 2011;21:243–58.
[38] Macpherson GJ, Breusch SJ. Metal-on-metal hip resurfacing: a critical review. Arch Orthop Trauma Surg 2011;131:101–10.
[39] Vendittoli P, Lavigne M, Roy A-G, Lusignan D. A prospective randomized clinical trial comparing metal-on-metal total hip arthroplasty and metal-on-metal total hip resurfacing in patients less than 65 years old. Hip Int 2006;16(Suppl. 4):73–81.
[40] Sawtooth Software. The CBC System for Choice-Based Conjoint Analysis (Technical Paper Series). Sawtooth Software Inc, Utah, 2008:1–26.
[41] Bridges JFP, Buttorff C, Groothuis-Oudshoorn K. Estimating Patients' Preferences for Medical Devices: Does the Number of Profile in Choice Experiments Matter? National Bureau of Economic Research, Cambridge, 2011.
[42] Train KE. Discrete Choice Methods with Simulation. Cambridge: Cambridge University Press, 2003.
[43] Louviere JJ, Flynn TN, Carson RT. Discrete choice experiments are not conjoint analysis. J Choice Model 2010;3:57–72.
[44] Kragt ME, Bennett JW. Attribute framing in choice experiments: how do attribute level descriptions affect value estimates? Environ Resour Econ 2011;51:43–59.
[45] Weintraub WS, Cohen DJ. The limits of cost-effectiveness analysis. Circ Cardiovasc Qual Outcomes 2009;2:55–8.