ORIGINAL RESEARCH ARTICLE

Comparing Outcomes of a Discrete Choice Experiment and Case 2 Best-Worst Scaling: An Application to Neuromuscular Disease Treatment

Vikas Soekhai[1,2,3] · Bas Donkers[1,4] · Jennifer Viberg Johansson[5,6] · Cecilia Jimenez-Moreno[7,8] · Cathy Anne Pinto[9] · G. Ardine de Wit[10] · Esther de Bekker-Grob[1,2]

Abstract

Background and Objectives Case 2 best-worst scaling (BWS-2) is an increasingly popular method to elicit patient preferences. Because BWS-2 potentially has a lower cognitive burden compared with discrete choice experiments, the aim of this study was to compare treatment preference weights and relative importance scores.

Methods Patients with neuromuscular diseases completed an online survey at two different moments in time, completing one method per occasion. Patients were randomly assigned to either first a discrete choice experiment or BWS-2. Attributes included: muscle strength, energy endurance, balance, cognition, chance of blurry vision, and chance of liver damage. Multinomial logit was used to calculate overall relative importance scores and latent class logit was used to estimate heterogeneous preference weights and to calculate the relative importance scores of the attributes for each latent class.

Results A total of 140 patients were included for analyses. Overall relative importance scores showed differences in attribute importance rankings between a discrete choice experiment and BWS-2. Latent class analyses indicated three latent classes for both methods, with a specific class in both the discrete choice experiment and BWS-2 in which (avoiding) liver damage was the most important attribute. Ex-post analyses showed that classes differed in sex, age, level of education, and disease status. The discrete choice experiment was easier to understand compared with BWS-2.

Conclusions This study showed that using a discrete choice experiment and BWS-2 leads to different outcomes, both in preference weights as well as in relative importance scores, which might have been caused by the different framing of risks in BWS-2. However, a latent class analysis revealed similar latent classes between methods. Careful consideration about method selection is required, while keeping the specific decision context in mind and pilot testing the methods.

Key Points for Decision Makers

Comparing case 2 best-worst scaling and discrete choice experiment outcomes within patients with neuromuscular diseases showed differences in relative importance scores but also comparable preference classes between the two methods.

Careful consideration when selecting either a discrete choice experiment or case 2 best-worst scaling to elicit patient preferences is necessary as these preferences may differ and the method should match the decision context.

✉ Vikas Soekhai
  soekhai@eshpm.eur.nl; vr.soekhai@minvws.nl

Extended author information available on the last page of the article

1 Introduction

There is an emerging consensus that patient preferences should be incorporated within decisions in the medical product lifecycle [1–4]. These preferences have become more important over time for the companies that develop new medical products and for the authorities that assess, regulate, and decide which products are effective, safe, well tolerated, and cost effective [5]. Yet, there are still outstanding questions related to which preference methods are best suited for each decision context and there are many different methods that can be used to gain insights into preferences. Studies by for example the Medical Device Innovation Consortium [6] and Soekhai et al. [7] provide an overview of several stated preference methods to elicit these preferences within the medical product life-cycle context.

△ Adis

One of the stated preference methods that has become increasingly popular to elicit patient preferences is best-worst scaling (BWS) [8, 9]. Best-worst scaling was introduced to obtain more preference information than a discrete choice experiment (DCE) by asking individuals not only to select their best but also their worst option, without a large increase in the cognitive burden of the elicitation task [8]. The literature distinguishes between three types of BWS: object case (case 1 BWS) where attributes (characteristics) are selected as best and worst, profile case (case 2 BWS) where attribute levels (values of characteristics) are selected as best and worst, and multi-profile case (case 3 BWS) where profiles are selected as best and worst [10]. For more details regarding BWS, see Louviere et al. [10] Case 2 BWS (hereafter: BWS-2) has received much attention in the preference literature, as this method is able to uncover attribute level importance, might reduce cognitive burden of the elicitation task by focusing on one profile at a time, and is relatively easy to design [11, 12].

Although BWS-2 is being used more frequently in health preference research, it cannot yet match the years of experience and the resulting body of work of DCEs in health preference research [13, 14]. In DCEs, respondents are presented with multiple-choice tasks including two or more hypothetical alternatives. These alternatives consist of a fixed set of attributes with varying attribute levels between the alternatives and choice tasks. Respondents are then asked to select their preferred alternative in each choice task. For more information about DCEs, see Hensher et al. [15] and Train [16].

There are few studies investigating differences between DCE and BWS-2 preference study outcomes. Studies from van Dijk et al. (hip replacement surgery) [11], Potoglou et al. (social care preferences) [17], and Severin et al. (priority setting for genetic testing) [18] are examples in which DCE and BWS-2 preferences have been compared. The aim of this study is to compare preference weights and relative importance scores obtained from both methods. In this study, we focused on treatment preferences for patients with neuromuscular diseases (NMD), which are rare diseases and often affect the central nervous system leading to impaired or reduced cognitive functioning [19–22]. General cognitive deficits have been described in over 60–70% of patients and the prevalence and severity depend on the age at onset of the disease. With an earlier onset of disease, the cognitive limitations are generally more severe than observed for adult phenotypes, which are classified as those with symptoms first diagnosed ≥ 20 years of age [23]. Comparing DCE and BWS outcomes in this study context is of interest, as DCEs generally require larger sample sizes, which is challenging for rare disease applications, and patients with NMD may have reduced cognitive functioning and the perception is that BWS-2

presents a lower cognitive burden for patients [24]. The latter is related to the fact that previous research showed that BWS-2 requires that all attributes are framed either positively or negatively (i.e., mixing benefits and risks leads to identifcation problems) [25], while in DCEs combining positive and negative attributes within one choice task is possible, making it cognitively more demanding. One of the aims of our study was to compare a DCE to a BWS case that is able to uncover attribute level importance (as the aim was to compare with DCE results), while at the same time reducing cognitive burden by focusing on one profile at a time (as lowering the cognitive burden for patients with NMD is important) and is relatively easy to design because no specific software is needed (important in clinical settings when eliciting preferences).

2 Methods

2.1 Study Population

A sample of adult patients with NMD was selected between May and December 2020. Respondents were mostly recruited through patient organizations and patient registries in the UK, USA, Canada, Australia, and New Zealand via e-mail, advertisements, and newsletters. Informed consent was obtained before the start of the survey. Respondents were included if they were 18 years of age or older, were self-reported as diagnosed with NMD with late onset (established diagnosis or first reported symptoms on or after 20 years of age), and had an active e-mail account to register. Respondents were excluded if they were unable to provide informed consent, complete the online survey, or with a reported history of encephalopathy or dementia (as these may have an impact on cognitive skills and ability to complete the survey). This study was approved by the Newcastle University Ethics Committee (Reference: 8840/2018).

2.2 Attributes and Attribute Levels

Potentially relevant attributes and attribute levels for a hypothetical medicinal treatment for patients with NMD were selected using a qualitative study for both DCE and BWS-2. The qualitative study included 52 participants who completed in-person semi-structured interviews or participated in focus group discussions. When designing the survey instruments, additional evidence such as a literature review and experience-based opinions from the key members of the team (patients, clinical experts, and methodological experts) were also considered. More details regarding these qualitative findings were reported elsewhere [26, 27]. These findings showed that 11 attributes were eventually narrowed down to six final attributes that were included in the DCE

Table 1 Attributes and levels for eliciting preferences with discrete choice experiment and case 2 best-worst scaling (including priors for discrete choice experiment design)

Attributes	Levels		
Muscle strength	Stays the same [reference category]	Improved by half [0.05, 0.15][a] [0.89, 0.45][b]	Cured [0.15, 0.25][a] [0.95, 0.49][b]
Energy endurance	Stays the same [reference category]	Improved by half [0.05, 0.15][a] [0.60, 0.30][b]	Cured [0.15, 0.25][a] [0.70, 0.36][b]
Balance	Stays the same [reference category]	Improved by half [0.05, 0.15][a] [0.42, 0.21][b]	Cured [0.15, 0.25][a] [1.05, 0.54][b]
Cognition	Stays the same [reference category]	Improved by half [0.05, 0.15][a] [0.05, 0.61][c]	Cured [0.15, 0.25][a] [0.15, 0.71][c]
Chance of (temporary) blurry vision	1% [reference category]	15% [−0.15, −0.05][a] [−0.59, −0.05][c]	30% [−0.25, −0.15][a] [−0.85, −0.15][b]
Chance of (permanent) liver damage	1% [reference category]	15% [−0.15, −0.05][a] [−0.65, 0.33][b]	30% [−0.25, −0.15][a] [−1.86, 0.95][b]

[a]Uniformly distributed pilot prior: minimum, maximum

[b]Normally distributed post-pilot updated prior: mean, standard deviation

[c]Uniformly distributed post-pilot updated prior: minimum, maximum

and BWS-2 as minimizing the cognitive burden was key: muscle strength, energy endurance, balance, cognition, chance of (temporary) blurry vision, and chance of (permanent) liver damage. Table 1 presents the attributes and attribute levels for DCE and BWS-2.

2.3 Design of DCE Choice Tasks

A Bayesian D-efficient design was generated for the DCE, in which the D-efficiency was maximized using Ngene software (Ngene, version 1.2.1) [28]. Pilot data from the first 51 respondents were used to update priors and their specific distribution[1] (see Table 1) as well as for further optimization of the design [28, 29]. The final DCE design used for the survey included 24 unique choice tasks, which were blocked into two blocks with 12 choice tasks each to reduce cognitive burden for respondents. The alternatives in each choice task were unlabeled and the attribute order was kept constant across all tasks [30].

After we collected data for 51 patients, we estimated a multinomial logit (MNL) model using the DCE data in order to update our priors to generate a more efficient design. We used a dummy specification and our analysis showed that for the attributes muscle strength, energy endurance, balance, and liver damage, attribute levels had the expected size, sign, and were statistically significant. For these attributes, we generated a new experimental design with a normal distribution with the estimation coefficient as mean and standard deviation = estimated coefficient/1.96 to account for preference heterogeneity. For the attributes cognition and blurry vision, the estimates were not as expected and we therefore decided to use the original experimental design choices for these two attributes.

2.4 Design of BWS Choice Tasks

For designing the BWS-2 choice tasks, an orthogonal main effect plan experimental design was used. An orthogonal main effect plan enables the independent estimation of preference weights for each attribute level [10]. Based on the number of attributes and levels, the orthogonal main effect plan indicated 18 choice tasks to be included in the experiment [31]. As the combination of negative and positive attributes in BWS-2 choice tasks can lead to identification problems, negative attributes (i.e., chance of blurry vision and chance of liver damage) were framed positively [25]. This means that for these attributes, attribute levels in Table 1 for BWS-2 included a 70%, 85%, and 99% chance

---

[1] A MNL model using the DCE data was estimated in order to update our priors to generate a more efficient design. We used a dummy specification and our analysis showed that for the attributes muscle strength, energy endurance, balance, and liver damage, attribute levels had the expected size, sign, and were statistically significant. For these attributes, we generated a new experimental design with a normal distribution with the estimation coefficient as mean and standard deviation = estimated coefficient/1.96 to account for preference heterogeneity. For the attributes cognition and blurry vision, the estimates were not as expected and we therefore decided to use the original experimental design choices for these two attributes.

of not experiencing blurry vision or liver damage. Attribute order was kept constant across all tasks.

2.5 Survey Design

The survey consisted of several sections. At T = 1 (first measurement with first part of the survey), this included (1) background questions such as demographics (age, sex, school or work situation, country of origin), recruitment platform, clinical characteristics (diagnosis and age of diagnosis), disease status, and a list of 18 activities along with questions about whether these were possible for the patient; (2) a short video introducing the preference task with an explanation of all attributes and attribute levels, (3) either BWS-2 (18 choice tasks) or a DCE (12 choice tasks) [randomly allocated], and (4) evaluation questions about the ease of understanding and answering, and the usefulness of the video instructions. At T = 2 (second measurement with second part of the survey), a short video introduced the other preference method and follow-up questions were also included [26]. To minimize the cognitive burden, the first set of choice tasks (either a DCE or BWS-2) and the second set of choice tasks were administered at different timepoints, with a 2-week period in between. In BWS-2, respondents had to select their best and worst attribute level, while in the DCE, respondents were asked about their preferences by choosing between two alternatives. The survey was designed using Lighthouse Studio (Sawtooth Software, version 9.8.1X). Examples of DCE and BWS-2 choice tasks are shown in Figure 1.

2.6 Statistical Analysis

Statistical analyses were performed using data from respondents who completed both BWS-2 and DCE tasks (including respondents from pilot). Following guidance from the literature, as well as our interest in investigating preference heterogeneity, identifying different respondent groups, and model fit, a latent class (LC) model was estimated to analyze choice data for both DCE and BWS-2 [10, 15]. While the standard multinomial logit (MNL) model, used as a starting point within this study, assumes that all respondents have identical preferences, the LC model deals with preference heterogeneity by assuming, based on the choices respondents made, that there are a fixed number of different groups of respondents (i.e., LCs) [16]. Within each group in a traditional LC model, each individual has identical preferences.

With the LC model, the utility (U) of an alternative for each LC in both the DCE and BWS-2 can be modeled as a linear function of the specific attributes and levels, with

$$U = \sum_{k=1}^{A} \sum_{j=1}^{J_k} \beta_{k,j} X_{k,j} + \varepsilon,$$ (1)

where there are A attributes with attribute k having J_k attribute levels, with $X_{k,j}$ equal to one if the attribute level j of an attribute k is available in the presented profile, $\beta_{k,j}$ are the utility parameters for the jth levels of attribute k, and ε is the random error term representing the unexplained part of utility. The LC model was programmed using R version 4.0.0 (Apollo package, version 0.0.1) to estimate the utilities for both the DCE and BWS-2 data, as well as for the ex-post descriptive analyses to characterize the LCs [32, 33]. For the DCE and BWS-2, "muscle strength stays the same" was included as the reference level (fixed at zero). The DCE also required a reference level within each specific attribute. To create a clear interpretation of attribute levels (for the attributes, muscle strength, energy endurance, balance, and cognition), the least attractive attribute levels were used as the reference level. For the other attributes, the most attractive attribute levels were selected as the reference level. This means that for muscle strength, energy endurance, balance, and cognition, preference weights increase when the attribute level value increases, while for the chance of blurry vision and the chance of liver damage the preference weights decrease with increasing attribute levels. To facilitate the comparison between the DCE and BWS-2, the utility levels relative to the corresponding attribute reference level were also estimated for BWS-2. Relative importance scores (RIS) of attributes were calculated (based on an MNL estimation) by looking at the maximum utility differences between the best and worst attribute levels within each specific attribute and comparing those between the DCE and BWS-2, while outcomes from the evaluation questions for both methods were also analyzed.

3 Results

A total of 140 patients completed both the DCE and BWS-2 part of the survey. Responding patients were mostly female (65%) and the median age was 54 years (with a range of 23–76 years). The majority of patients completed a higher (45%) or vocational (34%) education. Most patients reported that they were able to walk without an assistive device (36%), followed by 26% of the patients reporting able to walk but relying on an assistive device. A relatively large group of patients (23%) also reported able to walk and run without an assistive device (Table 2).

Figure 2 shows the overall (based on MNL) RIS calculations for both DCE and BWS-2. For DCE, (avoiding) liver damage had the highest relative importance, followed by muscle strength, energy endurance, balance, cognition, and (avoiding) blurry vision. For BWS-2, a different pattern was observed. Muscle strength had the highest RIS value, followed by energy endurance, balance, liver damage, cognition, and blurry vision. Preferences

Figure 1 Example of case 2 best-worst scaling and discrete choice experiment choice tasks

Table 2 Sample characteristics

Characteristic	DCE and BWS-2
Respondents	140
Sex	
Female	91 (65%)
Age (years)	
Median (range)	54 (23–76)
Highest level of education	
No formal schooling	1 (1%)
Elementary	5 (4%)
Secondary	21 (15%)
Vocational	48 (34%)
Higher	63 (45%)
No answer	2 (1%)
Disease status	
Walk and run without assistive device	32 (23%)
Walk without assistive device	50 (36%)
Walk but rely on assistive device	36 (26%)
Walk but using wheelchair part-time	19 (14%)
Fully rely on wheelchair	3 (1%)

BWS-2 case 2 best-worst scaling, *DCE* discrete choice experiment

for improving the typical impairments of NMD were similar across methods, with generally a high preference to improve muscle strength, energy, and (to a somewhat lesser extent) balance. Accounting for preference heterogeneity with LC, Figure 3a, b illustrate the relative importance of each attribute for each LC. Given the sample size, statistical measures of fit and aiming for a meaningful interpretation of the LCs, a three-class model was superior for both the DCE and BWS-2. The DCE LCs in Figure 3a reveal a group of patients in whom avoiding liver damage is by far the most important attribute, while there are also patient groups where improvement of balance and energy endurance are most important. For BWS-2, there is a patient group in which muscle strength is most important, while there is—similar to DCE—a patient group in which liver damage is considered the most important attribute (Figure 3b).

Table 3 presents the estimated LC preference weights for both preference methods. Focusing on the magnitude of these weights, for DCE overall, the more attractive levels were preferred above the less attractive levels with most attribute levels being statistically significant. This is however not the case in DCE class 2, in which most attribute levels are not statistically significant and where the utility of a 15% chance of liver damage was larger than the utility of a 1% chance of liver damage. The largest patient class (47%) was the class of patients in which liver damage was the most important attribute (class 3). For BWS-2, Table 3 shows that most attribute levels were statistically significant.

Additionally, all the more attractive attribute levels were preferred above the less attractive attribute levels. The largest classes of patients were the classes in which energy endurance (42%, class 1) and liver damage (41%, class 3) were the most important attributes.

To characterize patients in the three different DCE and BWS-2 LCs, ex-post analyses were conducted (Table 4) by making use of the sample characteristics in Table 2 because extending our LC model with a class membership model failed to converge owing to the relatively small sample. These results show that DCE LCs differed in terms of the level of highest education, sex, and age: DCE LC 2 included the highest percentage of female patients (72%), who were the youngest (median age 47 years) and who had the highest level of education (96% completed vocational or higher education). For BWS-2, LC 2 was also different compared with other classes: this class included the highest percentage of female patients (74%), who were the oldest (median age 58 years) and who were relatively less impaired by their disease (74% indicated that they were able to walk). The ex-post analyses in Table 4 also highlighted that there was a high level of concordance between patients in a specific DCE class and patients in the same BWS-2 class. More specifically, patients in the DCE class in which balance was the most important attribute (class 1) and in which liver damage was the most important attribute (class 3), had the highest probability to also be in BWS-2 LC 1 (energy endurance most important) and LC 3 (liver damage most important), respectively. This was however not the case for LC 2.

Table 5 presents the results from the evaluation questions regarding DCE and BWS-2. The results show that there are no statistically significant differences between methods for evaluation questions about help with the survey, difficulty of answering questions, and if the descriptions were sufficient. However, statistically significant (chi-squared test, *p*-value 0.04 < 0.05) differences were found between the DCE and BWS-2 about the difficulty of understanding the questions. The percentage of patients who found DCE choice tasks easier to understand (74%) was greater than the percentage of patients who found BWS-2 choice tasks easier to understand (62%). In order to gain knowledge specifically of the understanding of DCE and BWS-2 questions by patients, we also performed an individual patient-level analysis. This meant data were analyzed from the same patients that both saw DCE and BWS-2 (or the other way around) and completed both sets of evaluation questions about understanding the questions (see Table 6). Table 6 shows that overall patients who completed either DCE first (44% + 50% = 94%) or BWS-2 first (26% + 47% = 73%) both evaluated DCE more often as being very easy or easy, compared to BWS-2 (31% + 31% = 62% when DCE was the first method and 16% + 47% = 63% when BWS-2 was the first method).
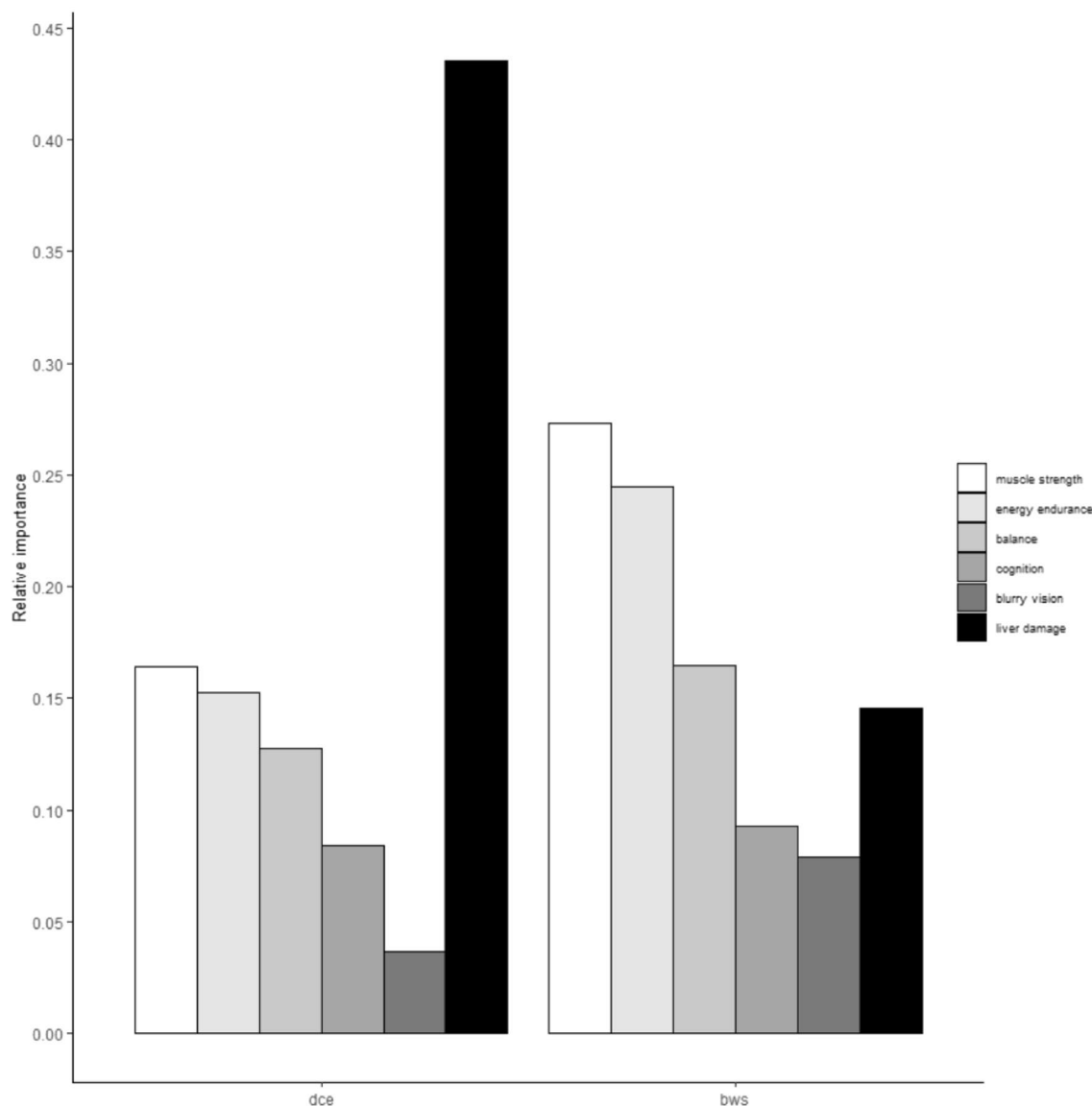
Figure 2 Overall relative importance score of attributes for discrete choice experiment (DCE) and case 2 best-worst scaling (BWS-2)

4 Discussion

In this study, preference weights and other outcomes (e.g., RIS) between DCE and BWS-2 were compared within patients with NMD. We conclude that the two methods lead to different preference weights as well as RIS values. However, accounting for preference heterogeneity, LC outcomes showed that patient classes look more similar, with a clear class of patients who both in DCE and BWS-2 indicated that liver damage was the most important attribute (class 3). For both preference methods, this class was among the largest class of patients. Additionally, patients that identified liver damage as most important (class 3) in DCEs also had the highest probability to be in the same class in BWS-2.

The ex-post analyses also showed that for both preference methods class 2 differed (which might be related to the small class size) in terms of descriptives (i.e., sex, age, education, disease status) compared with class 1 and class 3. Contrary to initial expectations, the proportion of patients who found DCE easier to understand was greater than the proportion of people who found BWS-2 easier to understand.

One of our main findings of this study was that both DCE and BWS-2 led to different outcomes. There are several stated preference studies comparing outcomes between these two methods. Studies by Van Dijk et al. [11], Potoglou et al. [17], and Severin et al. [18] showed similar outcomes between DCE and BWS-2. Differences between these studies and our study might first be related to differences in the

Figure 3 Relative importance of attributes for the discrete choice experiment (**a**) and case 2 best-worst scaling (**b**)
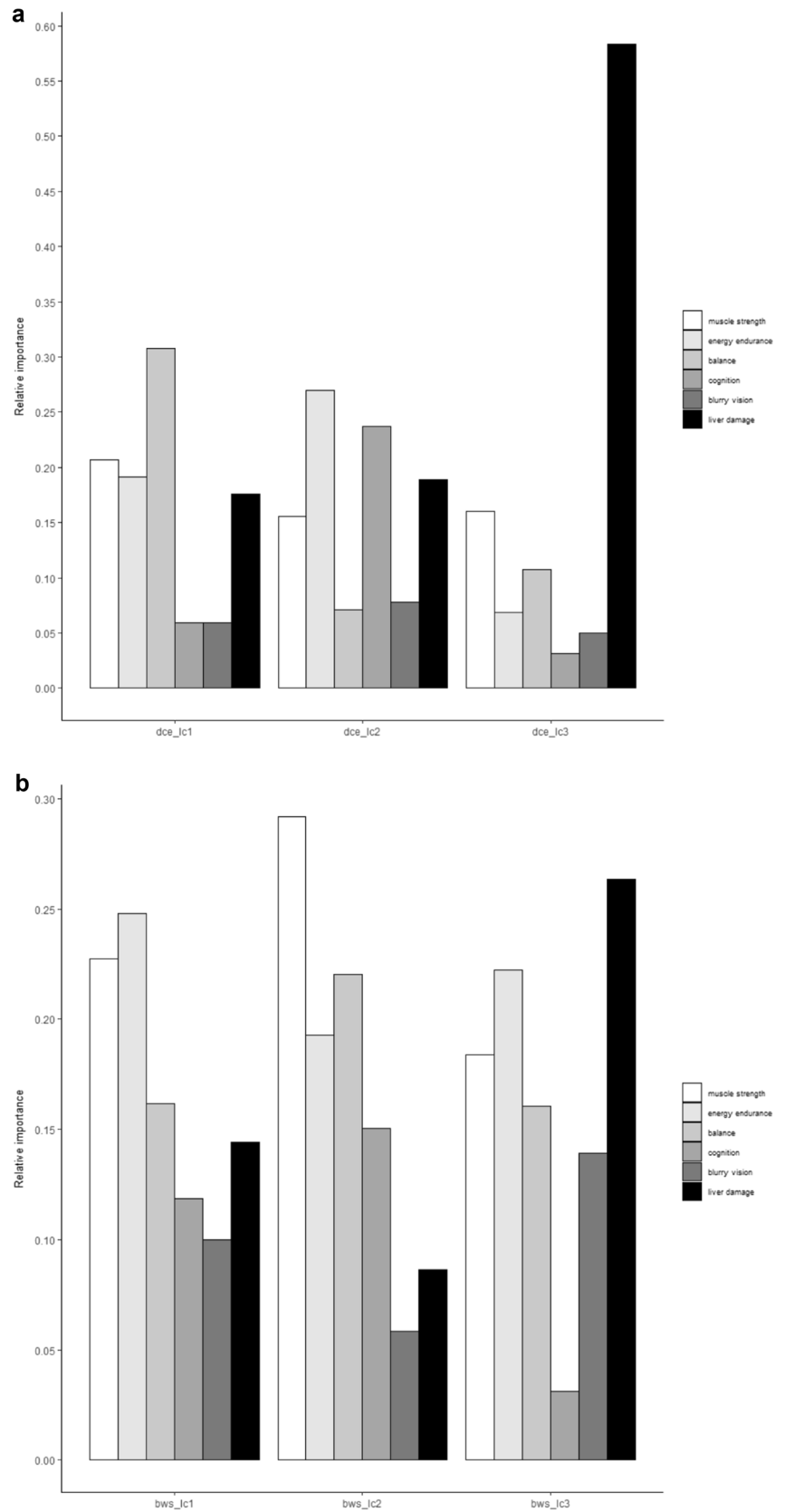
Table 3 Latent class analysis results for DCE and BWS-2

n = 140	DCE						BWS-2					
	Class 1		Class 2		Class 3		Class 1		Class 2		Class 3	
	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se
Muscle strength												
Stays the same	REF	–	REF	–	REF	–	REF	–	REF	–	REF	–
Improved by half	0.79	0.60	− 0.40	0.51	0.98**	0.27	2.25**	0.50	3.48**	0.74	2.15**	0.42
Cured	0.94**	0.38	0.54	0.46	1.80**	0.35	3.01**	0.52	4.83**	0.88	3.00**	0.46
Energy endurance												
Stays the same	REF	–	REF	–	REF	–	REF	–	REF	–	REF	–
Improved by half	0.62**	0.27	0.96	0.70	0.08	0.26	2.67**	0.47	1.85**	0.66	2.83**	0.38
Cured	0.87*	0.48	1.63	1.58	0.77**	0.25	3.28**	0.44	3.19**	0.45	3.63**	0.43
Balance												
Stays the same	REF	–	REF	–	REF	–	REF	–	REF	–	REF	–
Improved by half	1.02*	0.52	− 0.31	0.41	0.46	0.38	1.46**	0.41	2.52**	0.72	1.90**	0.35
Cured	1.40*	0.68	− 0.43	0.46	1.21**	0.41	2.14**	0.49	3.65**	1.13	2.62**	0.40
Cognition												
Stays the same	REF	–	REF	–	REF	–	REF	–	REF	–	REF	–
Improved by half	0.10	0.28	0.69*	0.39	0.05	0.20	1.24**	0.44	1.32**	0.54	0.28	0.34
Cured	0.27	0.19	1.43	1.17	0.35	0.23	1.57**	0.47	2.49**	0.85	0.51	0.39
Chance blurry vision (%)												
1	REF	–	REF	–	REF	–	REF	–	REF	–	REF	–
15	− 0.09	0.15	− 0.18	0.39	− 0.17	0.16	− 0.49**	0.24	− 0.16	0.13	− 1.39**	0.34
30	− 0.27*	0.15	− 0.47	0.37	− 0.56**	0.24	− 1.32**	0.33	− 0.97**	0.39	− 2.27**	0.43
Chance liver damage (%)												
1	REF	–	REF	–	REF	–	REF	–	REF	–	REF	-
15	− 0.51**	0.23	0.11	0.44	− 3.12**	0.43	− 1.81**	0.48	− 0.33	0.24	− 3.00**	0.44
30	− 0.80*	0.43	− 1.03	1.09	− 6.56**	0.86	−.45**	0.48	− 1.43**	0.51	− 4.30**	0.72
Reference levels[a]												
Muscle strength: stays the same	0		0		0		0		0		0	
Energy endurance: stays the same	0		0		0		− 0.21	0.30	0.69*	0.38	− 1.13**	0.41
Balance: stays the same	0		0		0		0.13	0.30	0.58	0.51	− 1.00**	0.44
Cognition: stays the same	0		0		0		0.64*	0.34	1.39**	0.57	− 0.20	0.37
Chance blurry vision (%): 1	0		0		0		1.19**	0.42	4.40**	0.86	− 0.65	0.58
Chance liver damage (%): 1	0		0		0		1.90**	0.41	5.60**	0.82	− 2.05**	0.71
class shares	0.35		0.18		0.47		0.42		0.17		0.41	
delta_class[b]	REF		− 0.66		0.29		REF		− 0.93		− 0.03	
beta_worst[c]	–						0.79					

Table 3 (continued)

n = 140	DCE						BWS-2					
	Class 1		Class 2		Class 3		Class 1		Class 2		Class 3	
	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se	β est.	Rob. Se
log likelihood	−876.46						−5788.99					

β est. β estimate, *BWS-2* case 2 best–worst scaling, *DCE* discrete choice experiment, *REF* reference, *Rob. Se* Robust standard error, *significant at 10%, **significant at 5%

[a]β estimate attribute levels estimated as an additional utility or disutility compared to the reference level

[b]delta_class parameters indicate the likelihood of being in a specific class compared to the reference class

[c]beta_worst parameter allows for scale differences between best and worst choices (hypothesis testing beta_worst = 1 showed no statistically significant outcomes)

health decision context. Working with different types of respondents and dealing with different types of decisions (e.g., treatment choice, priority setting) might lead to different behavior, different choices, and therefore different outcomes. Second, in our study, we explicitly framed negative attributes (i.e., blurry vision and liver damage) positively in BWS-2 choice tasks in order to avoid comparisons of positive and negative attributes with a BWS-2 choice task as this could lead to identification problems [25]. This was not the case in the previous studies. Additionally, there might also have been a framing effect in our study with regard to the attribute liver damage, as the word "permanent" was included in the choice task, which might be a reason why this attribute was being considered important in both DCE and BWS-2. For the other negative attribute in the DCE, risk of blurry vision, it was stated that problems would disappear once (hypothetical) medication was stopped. Indeed, this temporary negative side effect appeared to be far less important in patient decision making. However, although our study differs from some of the prior research studies comparing the two methods, our study outcomes are in line with a study by Whitty et al. [34] in which the authors also reported differences in relative preference weigths and preference orderings between DCE and BWS-2 in a priority setting context.

In our study, the same patient sample (*n* = 140) completed both 12 DCE and 18 BWS-2 choice tasks. Preference weights from LCs in Table 3 showed that especially in DCE LC 2, most attribute levels were not statistically significant (i.e., smaller t-values) compared with BWS-2. Furthermore, attribute levels in the DCE overall had smaller t-values compared with BWS-2. This can be an indication that given the same (small) sample size, BWS-2 might be the preferred method of choice when statistical power is important for decision making. It should be noted here that this can however only be concluded by assuming that the cognitive burden of the 12 DCE and 18 BWS-2 choice tasks are comparable. Our results also suggest a smaller utility scale for DCE, which suggests the need for larger sample sizes in a DCE compared with BWS-2, as also mentioned in previous work [24].

The BWS-2 literature states that one of the reasons BWS-2 could be an interesting preference method compared to a DCE is because of its lower cognitive burden [11, 12]. However, this study indicated that the proportion of patients who found the DCE easier to understand was greater than the proportion of patients who found BWS-2 easier to understand. It should be noted here that the number of choice tasks between DCE (12) and BWS-2 (18) was different and the lead-ins for DCE and BWS-2 tasks also differed because the pilot study showed that patients needed more guidance regarding the BWS-2 tasks, which may both have influenced the evaluation of the methods by patients. The findings in this study follow the trend as described in a study

Table 4 Ex-post analyses of latent class analysis of DCE and BWS-2

Characteristic	DCE			BWS-2			DCE and BWS-2
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Overall sample
Class share	35%	18%	47%	42%	17%	41%	–
Sex							
Female	65%	72%	63%	60%	74%	67%	65%
Age (years)							
Median (range)	54 (32–71)	47 (23–73)	55 (31–76)	54 (23–76)	58 (32–72)	53 (31–73)	54 (23–76)
Highest level of education							
No formal schooling	2%	0%	0%	0%	4%	0%	1%
Elementary	2%	0%	6%	3%	4%	3%	4%
Secondary	15%	4%	19%	21%	22%	7%	15%
Vocational	33%	40%	33%	29%	48%	34%	34%
Higher	44%	56%	42%	47%	22%	54%	45%
No answer	4%	0%	0%	0%	0%	2%	1%
Disease status							
Walk and run without assistive device	15%	24%	28%	19%	17%	30%	23%
Walk without assistive device	38%	40%	34%	32%	57%	31%	36%
Walk but rely on assistive device	33%	24%	21%	34%	9%	24%	26%
Walk but using wheelchair part-time	10%	12%	16%	12%	17%	14%	14%
Fully rely on wheelchair	4%	0%	1%	3%	0%	1%	1%
Crosstab	BWS-2 Class 1 (%)	BWS-2 Class 2 (%)	BWS-2 Class 3 (%)				
DCE class 1	42	21	38	–	–	–	–
DCE class 2	56	20	24	–	–	–	–
DCE class 3	37	12	51	–	–	–	–

BWS-2 case 2 best-worst scaling, *DCE* discrete choice experiment

by Himmler et al. [35] in which the authors found that DCE choice tasks were less cognitively burdensome than BWS-2 choice tasks. Whitty et al. [12] also reported that in their study the majority of respondents found it more difficult to complete BWS-2 compared with a DCE and most respondents preferred a DCE over BWS-2. The individual-level analysis also indicated that a DCE was more often evaluated as very easy or easy compared with BWS-2. However, these results should be interpreted with caution, as the sample ($n = 35$) of patients used for this analysis was very small. Therefore, the signal will expectedly contain quite some noise because of the low number of observations.

A strength of this study is that it is the first study focusing on differences in outcomes between DCE and BWS-2 with regard to a sample possibly hampered by cognitive limitations. As mentioned in the introduction, several studies have focussed on differences between DCE and BWS-2 outcomes. However, to our knowledge, there are no such studies conducted within the context of a sample with cognitive limitations specifically. This study is also important because NMD are considered rare diseases that often translate into relatively small sample sizes when

eliciting preferences. This study provides useful insights into how BWS-2 and DCE performed with a relatively small sample size.

At the same time, the relatively small sample size is a limitation of this study. In general, this will not be a problem when estimating choice models not accounting for preference heterogeneity (MNL). However, when estimating more sophisticated models like for example LC in this study, such small sample sizes could potentially lead to estimation problems. In this study, we were able to estimate an LC model, but the extension with a class membership model failed to converge. Therefore, descriptive ex-post analyses were conducted to characterize the different latent patient classes. Future studies should however focus on larger samples that have cognitive limitations to investigate preference heterogeneity more thoroughly. A further limitation of this study is the fact that no information about the exact cognitive limitations of patients was identified, analyzed, and accounted for. In order to get a better understanding about cognitive burden and using DCE or BWS-2, future studies should identify the cognitive limitations of patients. Another limitation is the fact that a different number of choice tasks for each patient

Table 5 Evaluation questions for DCE and BWS-2

Evaluation question	DCE ($n = 131$)[a]	BWS-2 ($n = 130$)[a]
Help with survey ($p = 1.00$)		
By myself	120 (92%)	119 (92%)
Some help	10 (7%)	10 (7%)
Someone else	1 (1%)	1 (1%)
Difficulty understanding questions ($p = 0.04$)		
Very easy	41 (31%)	27 (21%)
Easy	56 (43%)	53 (41%)
Not easy or difficult	28 (21%)	34 (26%)
Difficult	6 (5%)	11 (8%)
Very difficult	0 (0%)	5 (4%)
Difficulty answering questions ($p = 0.86$)		
Very easy	19 (15%)	18 (13%)
Easy	48 (37%)	41 (32%)
Not easy or difficult	40 (31%)	41 (32%)
Difficult	22 (16%)	28 (22%)
Very difficult	2 (1%)	2 (1%)
Description of benefits and risks was sufficient ($p = 0.41$)		
Yes	119 (91%)	114 (88%)
No	12 (9%)	16 (12%)

BWS-2 case 2 best-worst scaling, *DCE* discrete choice experiment

[a]Difference in the total number of patients in DCE and BWS-2 who completed the evaluation questions as these questions were not mandatory in the survey

Table 6 Individual-level evaluation questions for DCE and BWS-2

Evaluation question	DCE first method ($n = 16$)	
	DCE	BWS-2
Difficulty understanding questions		
Very easy	7 (44%)	5 (31%)
Easy	8 (50%)	5 (31%)
Not easy or difficult	0 (0%)	6 (38%)
Difficult	1 (6%)	0 (0%)
Very difficult	0 (0%)	0 (0%)

Evaluation question	BWS-2 first method ($n = 19$)	
	DCE	BWS-2
Difficulty understanding questions		
Very easy	5 (26%)	3 (16%)
Easy	9 (47%)	9 (47%)
Not easy or difficult	2 (11%)	3 (16%)
Difficult	3 (16%)	3 (16%)
Very difficult	0 (0%)	1 (5%)

BWS-2 case 2 best-worst scaling, *DCE* discrete choice experiment

was used in the DCE (12) and BWS-2 (18). This may have influenced the evaluation of the methods by patients. However, pilot testing showed that 18 choice tasks for BWS-2 was managable and given the number of attributes and levels, we were not able to create an experimental design in which the number of choice tasks between methods was equal. Future studies comparing these two methods should focus on an experimental design with an equal number of choice tasks for both methods.

5 Conclusions

This study showed that using either a DCE or BWS-2 leads to different preference weights as well as relative importance values. A potential reason lies in the way risks were framed (i.e., positive) in BWS-2, which was different than in a DCE. Patients indicated that DCE choice tasks were easier to understand than BWS-2 tasks. Accounting for preference heterogeneity, the LC analysis indicated comparable LCs in both the DCE and BWS-2, especially the class of patients that indicated that liver damage was the most important attribute. Hence, we advise careful consideration when selecting either BWS-2 or a DCE to elicit preferences as the results of this specific study suggest that BWS-2 is the preferred method of choice when dealing with small samples, while DCEs may be preferred when minimizing the cognitive burden is key and choice tasks include both benefits and risks. It will therefore be important that the method matches the size and characteristics of the patient population. Proper pilot testing in the target population will also be important. To support medical decision making, keep in mind the research and decision context will be key.

Declarations

Ethics approval Not available.

Consent to participate Not available.

Consent for publication Not available.

Availability of data and material The dataset is available upon request.

Code availability The software code is available upon request.

References

1. Hoos A, Anderson J, Boutin M, et al. Partnering with patients in the development and lifecycle of medicines: a call for action. Ther Innov Regul Sci. 2015;49(6):929–39. https://doi.org/10.1177/2168479015580384.

2. Anderson RM, Funnell MM. Patient empowerment: reflections on the challenge of fostering the adoption of a new paradigm. Patient Educ Couns. 2005;57(2):153–7. https://doi.org/10.1016/J.PEC.2004.05.008.

3. Smith MY, Hammad TA, Metcalf M, et al. Patient engagement at a tipping point: the need for cultural change across patient, sponsor, and regulator stakeholders: insights from the DIA Conference, "Patient Engagement in Benefit Risk Assessment Throughout the Life Cycle of Medical Products." Ther Innov Regul Sci. 2016;50(5):546–53. https://doi.org/10.1177/2168479016662902.

4. de Bekker-Grob EW, Berlin C, Levitan B, et al. Giving patients' preferences a voice in medical treatment life cycle: the PREFER Public-Private Project. Patient. 2017;10(3):263–6. https://doi.org/10.1007/s40271-017-0222-3.

5. US Food and Drug Administration. The voice of the patient: a series of reports from FDA's Patient-Focused Drug Development Initiative. 2017. https://www.fda.gov/industry/prescription-drug-user-fee-amendments/voice-patient-series-reports-fdas-patient-focused-drug-development-initiative. Accessed 8 Jun 2021.

6. MDIC. Medical Device Innovation Consortium (MDIC) patient centered benefit-risk project report: a framework for incorporating information on patient preferences regarding benefit and risk into regulatory assessments of new medical technology. http://mdic.org/wp-content/uploads/2015/05/MDIC_PCBR_Framework_Web1.pdf. Accessed 8 Jul 2018.

7. Soekhai V, Whichello C, Levitan B, et al. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discov Today. 2019;24(7):1324–31. https://doi.org/10.1016/j.drudis.2019.05.001.

8. 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(1):171–89. https://doi.org/10.1016/j.jhealeco.2006.04.002.

9. Mühlbacher AC, Kaczynski A, Zweifel P, Johnson FR. Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. Health Econ Rev. 2016;6(1):1–14. https://doi.org/10.1186/s13561-015-0079-x.

10. Louviere J, Flynn T, Marley AAJ. Best-Worst Scaling: Theory, Methods and Applications.Cambridge, U.K.: Cambridge University Press; 2015. https://doi.org/10.1017/CBO9781107337855.

11. van Dijk JD, Groothuis-Oudshoorn CGM, Marshall DA, IJzerman MJ. An empirical comparison of discrete choice experiment and best-worst scaling to estimate stakeholders' risk tolerance for hip replacement surgery. Value Health. 2016;19(4):316–22. https://doi.org/10.1016/j.jval.2015.12.020.

12. 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 Mak. 2014;34(5):638–54. https://doi.org/10.1177/0272989X14526640.

13. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics. 2014;32(9):883–902. https://doi.org/10.1007/s40273-014-0170-x.

14. Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics. 2019;37(2):201–26. https://doi.org/10.1007/s40273-018-0734-2.

15. Hensher DA, Rose JM, Greene WH. Applied choice analysis. 2nd ed. Cambridge: Cambridge University Press; 2015. https://doi.org/10.1007/9781316136232.

16. Train K. Discrete choice methods with simulation. 2nd ed. Cambridge: Cambridge University Press; 2009. https://doi.org/10.1017/CBO9780511805271.

17. Potoglou D, Burge P, Flynn T, et al. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. Soc Sci Med. 2011;72(10):1717–27. https://doi.org/10.1016/j.socscimed.2011.03.027.

18. 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(11):1202–8. https://doi.org/10.1038/ejhg.2013.36.

19. Weber YG, Roebling R, Kassubek J, et al. Comparative analysis of brain structure, metabolism, and cognition in myotonic dystrophy 1 and 2. Neurology. 2010;74(14):1108–17. https://doi.org/10.1212/WNL.0b013e3181d8c35f.

20. Lax NZ, Gorman GS, Turnbull DM. Review: central nervous system involvement in mitochondrial disease. Neuropathol Appl Neurobiol. 2017;43(2):102–18. https://doi.org/10.1111/nan.12333.

21. Johnson N, Imbrugia C, Dunn D, Duvall B, Butterfield R, Feldkamp M, Weiss R. Genetic prevalence of myotonic dystrophy type 1 (S23. 003). 2019.

22. Gorman GS, Schaefer AM, Ng Y, et al. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. Ann Neurol. 2015;77(5):753–9. https://doi.org/10.1002/ana.24362.

23. Meola G, Sansone V. Cerebral involvement in myotonic dystrophies. Muscle Nerve. 2007;36(3):294–306. https://doi.org/10.1002/mus.20800.

24. Flynn TN, Peters TJ, Coast J. Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data. J Choice Model. 2013;6:34–43. https://doi.org/10.1016/j.jocm.2013.04.004.

25. Soekhai V, Donkers B, de Bekker-Grob E. PNS295 best worst scaling: for good or for bad but not for both. Value Health. 2019;22:S813. https://doi.org/10.1016/j.jval.2019.09.2195.

26. Jimenez-Moreno AC, Pinto CA, Levitan B, Whichello C, Dyer C, Van Overbeeke E, de Bekker-Grob E, Smith I, Huys I, Viberg Johansson J, Adcock K, Bullock K, Soekhai V, Yuan Z, Lochmuller H, de Wit A, Gorman GS. A study protocol for quantifying patient preferences in neuromuscular disorders: a case study of the IMI PREFER Project. Wellcome Open Res. 2020;5:253. https://doi.org/10.12688/wellcomeopenres.16116.1.

27. Jimenez-Moreno AC, van Overbeeke E, Pinto CA, et al. Patient preferences in rare diseases: a qualitative study in neuromuscular disorders to inform a quantitative preference study. Patient. 2021;14(5):601–12. https://doi.org/10.1007/s40271-020-00482-z.

28. 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(1):3–13. https://doi.org/10.1016/j.jval.2012.08.2223.

29. Hensher, D., Rose, J., & Greene, W. (2015). Applied Choice Analysis (2nd ed.). Cambridge: Cambridge UniversityPress. https://doi.org/10.1017/CBO9781316136232.

30. De Bekker-Grob EW, Hol L, Donkers B, et al. Labeled versus unlabeled discrete choice experiments in health economics: an application to colorectal cancer screening. Value Health. 2010;13(2):315–23. https://doi.org/10.1111/j.1524-4733.2009.00670.x.

31. Hahn G, Shapiro S. A catalogue and computer program for the design and analysis of orthogonal symmetric and asymmetric fractional factorial designs. General Electric Research and Development Center; 1966.

32. Hess S, Palma D. *Apollo*: a flexible, powerful and customisable freeware package for choice model estimation and application. J Choice Model. 2019;32:100170. https://doi.org/10.1016/j.jocm.2019.100170.

33. Hess S, Palma D. Apollo: a flexible , powerful and customisable freeware package for choice model estimation and application. Stephane Hess & David Palma Choice Modelling Centre University of Leeds. Apollo user Man. 2019 April).

34. 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(4):e90635. https://doi.org/10.1371/journal.pone.0090635.

35. Himmler S, Soekhai V, van Exel J, Brouwer W. What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. J Choice Model. 2021;38:100265. https://doi.org/10.1016/j.jocm.2020.100265.

Authors and Affiliations

Vikas Soekhai[1,2,3] · Bas Donkers[1,4] · Jennifer Viberg Johansson[5,6] · Cecilia Jimenez-Moreno[7,8] · Cathy Anne Pinto[9] · G. Ardine de Wit[10] · Esther de Bekker-Grob[1,2]

[1] Erasmus Choice Modelling Centre, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

[2] Erasmus School of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

[3] Department of Public Health, Erasmus MC, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands

[4] Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

[5] Department of Public Health and Caring Sciences, Centre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden

[6] Institute of Futures Studies, Stockholm, Sweden

[7] Wellcome Centre for Mitochondrial Research, Newcastle University, Newcastle-Upon-Tyne, UK

[8] Patient Centered Research, Evidera, London, UK

[9] Pharmacoepidemiology, Merck & Co., Inc., Kenilworth, NJ, USA

[10] Juliuscenter for Healthsciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands