Journal of Choice Modelling 38 (2021) 100265

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: http://www.elsevier.com/locate/jocm

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

Sebastian Himmler [a,*], Vikas Soekhai [a,b], Job van Exel [a,c], Werner Brouwer [a,c]

[a] Erasmus School of Health Policy & Management, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands
[b] Department of Public Health, Erasmus MC, University Medical Center, P.O. Box 2040, 3000 CA, Rotterdam, the Netherlands
[c] Erasmus School of Economics, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Discrete choice experiment
Best-worst scaling
Cognitive burden
Response efficiency
Colour coding

ABSTRACT

To appropriately weight dimensions of quality of life instruments for health economic evaluations, population and patient preferences need to be elicited. Two commonly used elicitation methods for this purpose are discrete choice experiments (DCE) and case 2 best-worst scaling (BWS). These methods differ in terms of their cognitive burden, which is especially relevant when eliciting preferences among older people. Using a randomised experiment with respondents from an online panel, this paper examines the cognitive burden associated with colour-coded and level overlapped DCE, colour-coded BWS, and 'standard' BWS choice tasks in a complex health state valuation setting. Our sample included 469 individuals aged 65 and above. Based on both revealed and stated cognitive burden, we found that the DCE tasks were less cognitively burdensome than case 2 BWS. Colour coding case 2 BWS cannot be recommended as its effect on cognitive burden was less clear and the colour coding lead to undesired choice heuristics. Our results have implications for future health state valuations of complex quality of life instruments and at least serve as an example of assessing cognitive burden associated with different types of choice experiments.

1. Introduction

Developments like ageing populations and rapid advances in medical technology create challenges for budgets of publicly funded health care systems (de Meijer et al., 2013). Policy makers increasingly have to decide about which health care services to include in the basic benefits package, which should only be made available to certain subpopulations, and which should not be funded at all. Health technology assessment (HTA) generates valuable insights to support this decision-making process, using tools like cost-utility analysis. There, the benefits of health technologies are typically expressed in the incremental amount of health changes they produce. This is calculated based on data from generic, multidimensional quality of life instruments, and a weighting algorithm for the levels of the dimensions based on population or patient preferences (Neumann et al., 2016). Given that health and social care, for instance aimed at older persons, may affect more than health-related quality of life alone, more recently, broader well-being measure have been

* Corresponding author.
E-mail addresses: himmler@eshpm.eur.nl (S. Himmler), soekhai@eshpm.eur.nl (V. Soekhai), vanexel@eshpm.eur.nl (J. van Exel), brouwer@eshpm.eur.nl (W. Brouwer).

https://doi.org/10.1016/j.jocm.2020.100265
Received 30 April 2020; Received in revised form 17 November 2020; Accepted 19 November 2020
Available online 1 December 2020

developed (Makai et al., 2014). These could facilitate cost-utility analyses with a broader scope in terms of relevant outcomes but require obtaining preferences for different 'well-being states' ideally anchored on death.

The measurement of population and patient preferences in health care is a rapidly developing field, with a plethora of qualitative and quantitative methods to the disposal of researchers and practitioners (Soekhai et al., 2019). One of the most popular methods over the last decade was the discrete choice experiment (DCE). Increasingly, population and patient preferences in health care are obtained using DCEs (Soekhai et al., 2019). The 'standard' DCE entails asking respondents to choose between two or more alternatives (Ryan et al., 2008) and is widely used for weighting quality of life instruments (Mulhern et al., 2018).[1] Another preference elicitation approach that gained traction over the last years also in this context, is best-worst scaling (BWS). There are three different forms of BWS – object case, profile case, and multi-profile case. The following will focus on profile case, or also called case 2 BWS, where individuals have to select a best and a worst option from a list of dimension levels or items (Flynn and Marley, 2014). Case 2 BWS was applied to value different quality of life instruments before (Cheung et al., 2016). This includes the ICECAP-O, a well-being measure specifically aimed at older people (Coast et al., 2008).

While both DCE and BWS provide numerical estimates of the relative importance of the different levels and dimensions of the respective quality of life or well-being instrument, previous research directly comparing DCE and BWS has shown that the choice between these approaches is not neutral as resulting preference estimates can differ (see e.g. Krucien et al., 2017). According to a recent review comparing DCE and BWS, there seems to be no conclusive evidence yet on which of the methods should be preferred in terms of the validity of the estimates (Whitty and Oliveira Gonçalves, 2018). Both methods assume different choice processes and ultimately may be seen to answer more or less subtly different questions. Some researchers prefer DCEs because the modelled choice processes have a strong theoretical foundation in random utility theory (Louviere, 2004). Providing choices between multiple alternative profiles can also be considered as a more realistic way of the decision-making process compared to selecting a best and worst option from a list of items. Another advantage of DCEs in the context of health state valuation, is that utilities can more easily anchored onto the full health (or well-being)-dead scale. On the other hand, some argue that profile case BWS is to be preferred as it is a more efficient way of collecting data compared to DCE since each task entails two choices. Moreover, cognitive burden of BWS tasks may be lower, since individuals only need to focus on one set of attributes and levels in each choice task, compared to multiple in DCEs. Some specifically claim that it would be recommendable to choose case 2 BWS if DCE tasks are considered to be too burdensome (Flynn, 2010; Potoglou et al., 2011). However, Whitty and Oliveira Gonçalves (2018) conclude that there is no clear evidence for an advantage of BWS regarding participant acceptability in terms of feasibility of administration or response efficiency. The response efficiency, that is, the cognitive burden associated with choice tasks, is important as it influences choice consistency, respondent fatigue and the use of simplifying choice heuristics (Jonker et al., 2019), which could subsequently influence the validity of the preference estimates.

Due to the ageing of the population, the need for economic evaluations of health and social care services targeted at older people can be expected to increase. This makes accurately measuring and weighting quality of life dimensions in this population very important, and choosing the appropriate methodology to do so, all the more relevant. If one decides, as we do here, that an instrument aimed at older people should be weighted using older peoples' preferences,[2] one needs to be aware of an additional aspect: Since there is a large variation in the level of cognitive abilities within older people, the design of choice experiments for this population should especially be wary of the complexity and subsequent cognitive burden of the choice task format in order to enable obtaining valid and reliable responses (Milte et al., 2014). Measuring and weighting quality of life or well-being outcomes inaccurately may ultimately lead to sub-optimal policy recommendations for resource allocation to health or social care services aimed at older people.

Specific evidence about the cognitive burden of DCE and case 2 BWS for valuing quality of life measures among older people is lacking. Therefore, the main aim of this study was to assess the cognitive burden and incidence of simplifying choice heuristics in DCE and case 2 BWS choice tasks among older people in this context. Another aim was to test the impact of the use of colour coding on the cognitive burden and choice behaviour of case 2 BWS tasks, which has been assessed for DCEs before (Jonker et al., 2019).

2. Methods

We set up a randomised experiment with three study arms to examine the cognitive burden and choice behaviour attached to three respective choice task formats for valuing a quality of life instrument: a colour coded and level overlapped DCE (5 out of 9 dimensions), a case 2 BWS and a colour coded case 2 BWS.[3] In the applied colour coding, five shades of one colour correspond to the five levels of attributes of the used instrument, with darker shades representing the least desirable levels. The rationale behind this type of coding in the DCE is that it helps respondents to identify differences between the alternatives, and higher and lower levels, while not nudging respondents to only focus on the differing attributes, what e.g. exclusively highlighting the non-overlapped levels would do, or introducing strong prejudgments on the severity of the levels by using e.g. a traffic light colour coding.

We chose an online setting with participants from an online panel for our study, as this administration and sampling mode facilitates reaching a sufficiently large number of respondents for health state valuation studies, which is also why it is used in most such studies by now (Mulhern et al., 2018).

The quality of life measure used in the experiment was the recently developed Well-being of Older People instrument (WOOP)

[1] This is also known as 'health state valuation'.

[2] Whose values to elicit is debatable in health state valuation in general. We decided to use older peoples' preferences as the WOOP is intended to inform allocation decisions only within care for older people.

[3] In the remainder of the paper, BWS refers to case 2 BWS.

(Hackert et al., 2019). Examining the cognitive burden of a valuation task is especially important in the context of this new instrument for measuring the general/overall quality of life of older people: First, the WOOP consists of nine dimensions with five levels each, which requires complex choice tasks. Second, as preferences should be based on an older population, cognitive burden is of special relevance. The profiles shown to respondents in both DCE and BWS tasks corresponded to well-being states, described using the nine dimensions of the WOOP (i.e. physical health, mental health, social life, receive support, acceptance and resilience, feeling useful, independence, making ends meet, living situation).[4] In designing the choice tasks and their visual representation, we followed methodological work on the use of colour coding and level overlap in DCEs aimed to reduce task complexity (Jonker et al, 2018, 2019; Maddala et al., 2003). To enable a more direct comparison and to test the impact of colour coding on task complexity in BWS, which has not been studied before, the randomised experiment included a colour coded BWS and a regular BWS.

Important to note here is that the design was generated to test the cognitive burden and choice behaviour of older people, not to provide model estimates for the different methods. Due to the large descriptive system of the WOOP, this would have required estimation of 36 parameters in the DCE and 45 parameters in the BWS, a blocked design and a much larger sample size. While a comparison of model estimates would have been interesting, this was not our current research aim.

2.1. Survey structure and randomisation

The structure of the experimental survey is shown in Fig. 1. First, respondents were asked to complete the WOOP instrument to become familiar with its dimensions and levels. Afterwards, they were randomized 1:1:1 to the three study arms: colour coded DCE (1), colour coded BWS or BWSc (2), and regular BWS (3). The randomisation was preferred over having the same respondents completing both DCE and BWS tasks, to have avoid the different parts of the experiment influencing each other and to stay as close as possible to standard DCE and BWS experiments. Furthermore, two full sets of valuation tasks per respondents were considered to be too burdensome. Respondents were familiarized with the presentation of well-being states in the subsequent experiment by showing them their own profile in DCE or BWS format based on the answers they previously gave to the WOOP instrument. The choice task formats were introduced by a simple DCE or BWS task, where participants had to select between two types of fruits or chose the best and worst type of fruit from a list. The second part of the warm-up comprised of a choice task, as used in the subsequent experiment, providing further instructions. Subsequently, a block of six choice tasks was administered, followed by two simple break questions on an unrelated topic to interrupt the monotony and reduce respondent fatigue of answering the choice tasks. Then, a second block containing seven tasks concluded the randomized part of the questionnaire, leading to a total of 13 choice tasks per respondent. All respondents subsequently had to fill in three blocks of evaluation questions on a 5-point Likert scale, before providing some sociodemographic information at the end of the survey.

2.2. Survey administration and participants

The survey was programmed using Sawtooth software version 9.7.2 (Sequim, WA). We used Prolific.co to recruit survey participants, a platform for online subject recruitment specifically for research purposes (Palan and Schitter, 2018). Given our aim to assess the cognitive burden of the choice tasks in a sample of older people, being aged 65 or above was used as inclusion criteria (which is also the target population of the WOOP). Since this age group was underrepresented in the online panel, we had to combine respondents from the two largest country panels of Prolific.co, UK and U.S. residents, to obtain a reasonably sized sample. At the time of data collection, in October 2019, the potential respondent pool contained around 1,000 individuals. Using quota sampling, we aimed for 150 respondents for each of the three study arms. Respondents received a monetary compensation for participating, which was oriented on the mean completion time and averaged to an aggregated hourly reward of £7.62. To test the functionality of the survey and whether respondents understood the choice tasks, six think-aloud interviews with UK residents aged 65 and above were conducted (two per study arm) prior to the main data collection. These interviews showed that participants understood and appropriately engaged in the choice tasks (i.e. traded-off or considered multiple items).

2.3. Experimental design of DCE and BWS

Attributes and levels in the DCE and items of the BWS were based on the dimensions and levels of the WOOP instrument (Appendix A). This created a rather complex DCE setup with nine dimensions with five levels each and a BWS instrument with 45 items. WOOP well-being states were consequently defined by selecting one of the five levels from each of the nine dimensions for both DCE and BWS. In the DCE, respondents were repeatedly presented with two well-being states and asked to indicate, which of the two they preferred. An opt-out option was not included as this is uncommon in DCEs for health state valuation (Mulhern et al., 2018). In the BWS, a list of nine well-being items corresponding to one well-being state was shown to respondents. Participants then had to select the aspect that they most preferred (best) and the aspect that they least preferred (worst). 'Most' and 'least' is one of the options that are used for describing a best and worst choice (Huynh et al., 2017).[5]

To ensure that the choice tasks had a similar level of complexity compared to a regular choice experiment, choice tasks were created

[4] Appendix A contains an updated version of the full descriptive system of the WOOP, with some formulations differing slightly compared to the version used in this study.

[5] 'Most' and 'least' may have a slightly different interpretation than 'best' and 'worst', but this should not have an impact on cognitive burden.

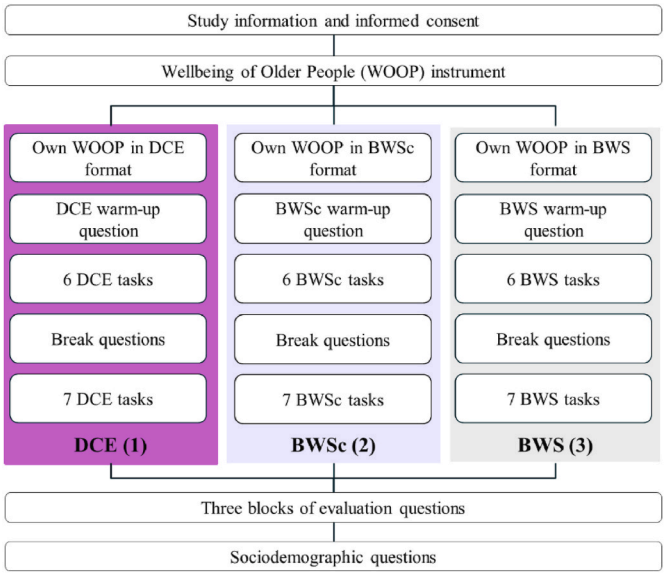

Fig. 1. Survey structure and experimental arms.

using standard design methodology as outlined in the subsequent paragraph. The literature on health related DCEs specifically targeted at older people was reviewed (in total 22 papers were studied) to inform the number of choice tasks. The number of choice tasks per respondent varied between 6 and 16 with a mean of 9.2. We opted to select a number of choice tasks at the upper end of this range (13) to capture fatigue effects (examples of this literature are e.g. Arendts et al., 2017; Franco et al., 2016; Milte et al., 2014) and because we anticipated this might be close to the approximate number in the actual valuation study of the WOOP. The 13 choice tasks consisted of 10 DCE choice tasks, two that repeat one of them, and one choice task to test for dominance.

The ten DCE choice were selected with help of Ngene design software (Version 1.2.1). To accommodate for level overlap (five out of the nine dimensions), which has been shown to reduce task complexity by Maddala et al. (2003) and Jonker et al. (2018), Ngene required a dataset including all possible candidate sets, i.e. combinations of two health states with five overlapped levels. To pragmatically reduce this to a feasible number, 5,000 out of the 1,953,125 possible health states were randomly selected and combined in MATLAB (MathWorks). Out of the obtained 25 million possible sets, we excluded the ones without the specified amount of overlap and randomly selected 1,000 sets out of the remaining 386,030 overlapped sets. Ngene was then used to select 10 choice tasks out of the 1,000 candidate sets by optimizing for a conditional logit, main effects model (Appendix C contains the utility function) with 36 parameters corresponding to four of the five levels of each of the nine dimensions of the WOOP instrument. Small priors ranging from 0 to −0.25 were assumed, following the logical ordering of the WOOP levels. Besides the think-aloud interviews no further pilot testing was conducted.

An orthogonal main effects plan using Sawtooth software version 9.7.2 (Sequim, WA) was applied to generate 1,000 blocks of 10 choice tasks for the BWS experiment. Multiple levels from the same WOOP dimension were prohibited to appear in the same task. Following Flynn et al. (2015), to prevent uninformative sets, we reduced the occurrences of tasks with either only one top or bottom WOOP level by deleting all versions where this occurred more than 3 times in the 10 tasks. Out of the remaining 78 versions, one version was randomly selected to be used in the experiment.

We selected one of the created DCE and BWS choice tasks to appear as the second choice task and repeated the tasks at position 8 and 13, to test choice consistency, adding two choice tasks to the original 10 created tasks. In order to reduce the amount of noise in the answers, we chose tasks, which were expected to have a certain degree of utility difference between profiles in the DCE arm or provided somewhat clear BWS choices (the repeated choice tasks are shown in Appendix B). When this task was repeated the second time, the intensity colour coding of the BWS task was intentionally reversed, to mislead respondents in order to assess the dependence on the colour codes. A dominant DCE choice task and a BWS task, which was expected to have a clear best and worst choice were additionally created and added at position 6 to test the attention level of respondents, adding a third and final choice task to the original ten created tasks.[6] The order of the dimensions (or attributes) was the same for all respondents within elicitation method and fixed for both DCE and BWS tasks to further reduce task complexity. The only difference in attribute order between DCE and BWS tasks was that physical and mental health attributes where positioned in the middle of the BWS tasks, as we anticipated that these would be important dimensions and wanted to avoid respondents making their best and worst choice merely on the top without going over the remaining items. All respondents received the same 13 DCE tasks in study arm 1. Respondents in study arms 2 and 3 received the same 13 BWS

[6] We decided against including results of this task in the final analysis, as such tasks are inherently difficult to compare between DCE and BWS (Whitty and Oliveira Gonçalves, 2018).

tasks.

2.4. Visual presentation of choice tasks

The general visual representation of the choice tasks followed current practice, with the exception that intensity colour coding was added to the choice tasks in study arms 1 and 2. Different shades of purple represented the different attribute levels, with the darker shades of purple highlighting the worse and the lighter shades and light blue expressing the better WOOP attribute levels in both the DCE and the colour coded BWS tasks. In the explanation of the colour coding in the survey, 'better levels' (e.g. very well able to cope, feeling very independent, no problems with physical health) were formulated as 'positive aspects' and 'worse levels' (e.g. barely able to cope, feeling very dependent, severe problems with physical health) as 'negative aspects' (e.g. Fig. 2). This type of colour coding was previously used for DCEs by Jonker et al. (2017, 2018, 2019) and was found to reduce task complexity as well as attribute non-attendance, and was especially effective in combination with attribute level overlap. It was also shown that colour-coding does not introduce bias in the choices and does not affect the relative importance of attributes (Jonker et al., 2019). The purple colour scheme was specifically designed to accommodate for the most prevalent forms of colour blindness. Additionally, shades of purple do not prompt natural or perceived value judgements, as opposed to for example traffic light colour coding.

Fig. 2 shows an example of the layout of the colour-coded (light blue to deep purple) and overlapped (five out of the nine dimensions) DCE choice task. Level descriptions of the WOOP instrument (Appendix A) were shortened for clarity, level labels were highlighted in bold, and attribute descriptions appeared merely as mouseovers on the attribute labels to reduce the amount of text. Fig. 3 shows examples of both colour coded and non-colour coded BWS tasks. Descriptions of attributes were also included as mouseovers, while the item text contained the full WOOP level descriptions.

2.5. Statistical analysis

To assess and compare the cognitive burden and possible choice heuristics associated with the three formats of choice tasks, three types of data were analysed. First, objective measures including mean choice task completion time, development of time per task (assessing learning effects) and drop-out rates were calculated and compared. Second, mean response scores of the three blocks of debriefing questions on perceived choice complexity, the number of choice tasks, and choice strategies used, were obtained. The latter aimed to identify the extent to which respondents engaged in simplifying choice heuristics. This included two statements relating to the number of attributes commonly considered during the choice tasks, also known as attribute non-attendance (Yao et al., 2015), and a statement on deciding that all attributes/dimensions are equally important. This statement implies that respondents merely count up the attribute level positions instead of trading-off attributes in the DCE, or focusing mostly on the level positions, irrespective of attribute, in the BWS format.

Third, revealed cognitive burden regarding choice consistency and (simplifying) choice behaviour was assessed based on the actual choices of respondents. This included calculating the proportion of respondents providing the same answers to the twice repeated choice task. For the BWS arm, a consistent response was defined as providing the same answer for either best or worst option, following

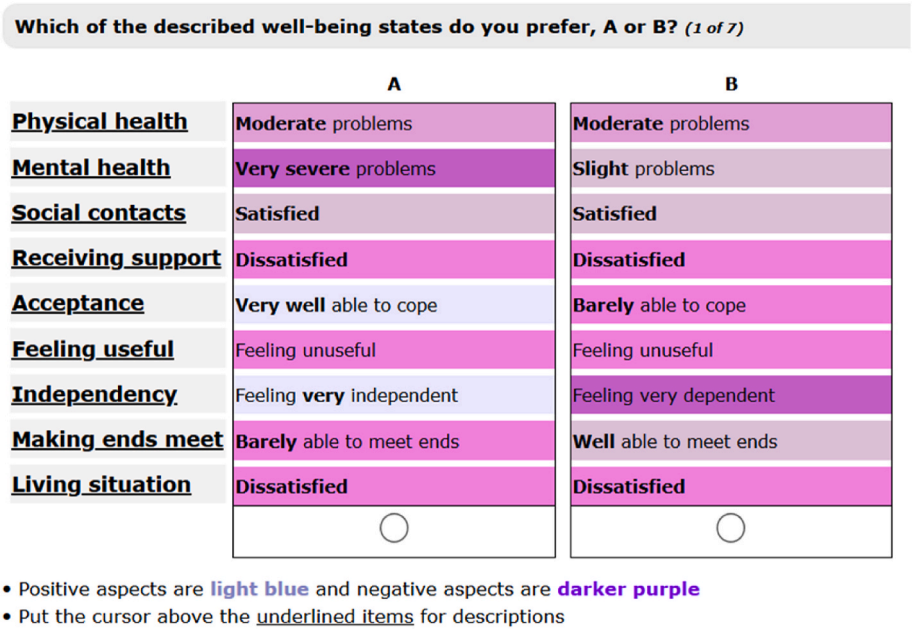

Fig. 2. Visual presentation of DCE choice task with colour coding and level overlap. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Imagine living in this well-being state and select which aspect you would most prefer, and which aspect you would least prefer. *(1 of 6)*

Most	Well-being state	Least
○	I am dissatisfied with my social contacts	○
○	I am reasonably satisfied with the support I receive	○
○	I am reasonable able to deal with my circumstances and changes therein	○
○	I feel reasonably useful	○
○	I have slight problems with my physical health	○
○	I have very severe problems with my mental health	○
○	I feel very dependent	○
○	I am very well able to make ends meet	○
○	I am dissatisfied with my living situation	○

- Positive aspects are **light blue** and negative aspects are **darker purple**
- Put the cursor above the underlined items for descriptions

Imagine living in this well-being state and select which aspect you would most prefer, and which aspect you would least prefer. *(1 of 6)*

Most	Well-being state	Least
○	I am dissatisfied with my social contacts	○
○	I am reasonably satisfied with the support I receive	○
○	I am reasonable able to deal with my circumstances and changes therein	○
○	I feel reasonably useful	○
○	I have slight problems with my physical health	○
○	I have very severe problems with my mental health	○
○	I feel very dependent	○
○	I am very well able to make ends meet	○
○	I am dissatisfied with my living situation	○

- Put the cursor above the underlined items for descriptions

Fig. 3. Visual presentation of colour-coded and non-colour-coded BWS choice task. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Krucien et al. (2017). Furthermore, we estimated a lexicographic score, which provides information on trading between attribute levels and dominant choice behaviour. This score was obtained also following an approach applied by Krucien et al. (2017): First, the proportion of choices based on one attribute on an individual level was calculated. Assuming respondents exhibit dominant preferences for an attribute given proportions above 90% (DCE) and 50% (BWS), the lexicographic score was obtained by calculating the proportion of respondents with such preferences.

To test the impact of colour coding on the choice behaviour and strategies in the BWS study arms, the shares of responses based on top and bottom levels of the WOOP dimensions were calculated. Additionally, results from the second repeated choice task, where the intensity colour coding was reversed, was used to assess the dependence on the colour scheme.

Statistical significance was assessed using Wilcoxon-rank sum tests for the Likert scale data (de Winter and Dodou, 2010) and chi-squared tests or Fisher exact tests for proportions. A significance level of 10% was used throughout the analysis. Stata 15 was used for all calculations.

3. Results

3.1. Sample characteristics, dropouts, and completion time

A total of 477 participants successfully started with the experiment and were randomly allocated to the three study arms. No respondent dropped out in study arm 1 (DCE). One of the three dropouts in study arm 2 (BWSc) occurred during the choice tasks and two afterwards. Of the five respondents dropping out in study arm 3 (BWS), four dropouts occurred during answering the BWS tasks and one at a later stage. Fisher exact tests indicated that the difference in total drop-out rates was significantly lower in study arm 1 compared to study arm 3 (0% vs. 3.2%, p-value = 0.029). The difference to study arm 2 was not significant (0% vs. 1.9%, p-value = 0.248).

The characteristics of the remaining sample, split by study arm, are shown in Table 1. The randomisation lead to well-balanced samples regarding most sociodemographic aspects, health status (EQ-5D-5L) and well-being (WOOP). 63.7% of the overall sample was younger than 70 years, 34.6% was aged between 70 and 79 years, and 1.7% were aged 80 years and above with 87 years as the maximum age observed.

The average time it took respondents to complete all 13 choice tasks was 6.0 min (SD 3.1) for the DCE tasks, 7.6 min for the colour coded BWS tasks (SD 4.9) and 7.2 min for the standard BWS tasks (SD 4.6). T-tests indicated that choice task completion time was significantly lower for the DCE tasks compared to the two sets of BWS tasks ($p < 0.001$ and $p = 0.007$). Fig. 4 plots the mean and median completion times for each choice task separated for each study arm. Differences were most pronounced in the beginning with choice task completion time following a downward trend, likely resulting from learning effects. Finding large differences in mean, but moderate in median answering time in the beginning indicates that some respondents found it particularly difficult to work with and understand the BWS question format compared to the DCE format. On aggregate, respondents in study arm 1 answered each choice task faster compared to the BWS study arms, except for one choice task. Differences within the two BWS study arms were less pronounced with the notable exception of choice task 13, where the intensity colour coding was reversed (e.g. light blue corresponded to the worst level and deep purple to the best).

3.2. Self-reported cognitive burden of tasks and number of choice tasks

Mean response scores of the three blocks of debriefing questions and results from significance tests comparing the mean scores across study arms are shown in Table 2. DCE choice tasks appeared to be superior in terms of clarity of the tasks and whether tasks were comprehensible from the beginning. Respondents found the presented states easier to image in the BWS tasks, which admittedly confronted participants only with one well-being state instead of two in the DCE. Colour coded BWS choice tasks were evaluated to be less clear then non-colour coded BWS tasks.

Results from the second block of questions indicated that participants from the DCE study arm found the number of choice tasks easier to manage, were more able to stay concentrated over all choice tasks, and could have answered more tasks, compared to the BWS study arms, with most differences being statistically significant. Colour coding the BWS tasks appeared to have a positive effect on the

Table 1
Main characteristics of analysis sample per study arm.

	DCE (1)	BWSc (2)	BWS (3)
Age in years	69.3	69.1	68.9
Female (%)	0.65	0.60	0.62
Years of education	16.1	15.8	15.8
Country of residence: UK (ref. U.S.) (%)	0.57	0.54	0.52
Employed (%)	0.33	0.29	0.28
EQ-5D-5L utilities (0–1)	0.83	0.82	0.82
WOOP (Sum score rescaled to 0–1)	0.81	0.79	0.82
Number of completes (N)	159	158	152

Note: EQ-5D-5L tariff from Devlin et al. (2018).

Fig. 4. Mean and median completion times per choice task within each study arm.

Table 2
Mean response score of cognitive debriefing questions.

Question on Likert scale from 1 to 5 (5 = strongly agree)	DCE (1)	BWSc (2)	BWS (3)
Self-reported cognitive burden			
The choice tasks were clear	4.45$^{\dagger ALL}$	4.11$^{\dagger ALL}$	4.25$^{\dagger ALL}$
I could easily choose between the alternatives	3.55	3.65	3.62
I fully understood the choice tasks from the beginning	4.75$^{\dagger ALL}$	4.26$^{\dagger 1}$	4.36$^{\dagger 1}$
The tasks got easier after answering several	3.77	3.87	3.84
I found some of the presented states difficult to imagine	3.43$^{\dagger 3}$	2.97$^{\dagger 1}$	2.84$^{\dagger 1}$
Number of choice tasks			
The number of choice tasks was manageable	4.64$^{\dagger 3}$	4.54	4.50$^{\dagger 1}$
It was difficult to stay concentrated over all choice tasks	1.72$^{\dagger 3}$	1.94	1.92$^{\dagger 1}$
I could have answered more choice tasks	4.07$^{\dagger ALL}$	3.91$^{\dagger ALL}$	3.66$^{\dagger ALL}$
Answering another block of six 6 choice tasks would be manageable	4.43$^{\dagger ALL}$	4.19$^{\dagger 1}$	4.18$^{\dagger 1}$
Choice strategies			
I compared all dimensions/items before making my choice	4.72	4.77	4.79
I decided all dimensions/items are equally important	2.86$^{\dagger 3}$	3.00	3.20$^{\dagger 1}$
I always used the same 1 or 2 well-being dimensions to make my choice	3.04$^{\dagger ALL}$	2.65$^{\dagger 1}$	2.57$^{\dagger 1}$

Note: † $p < 0.10$ of Wilcoxon rank-sum test comparing study arms 1, 2, and 3.

Table 3
Revealed choice behaviour.

	DCE (1)	BWSc (2)	BWS (3)
Non-trading or dominant choice behaviour			
Lexicographic score	28.9%[†ALL]	79.1%	80.1%
Choice consistency			
% failed a consistent response to repeated choice task (1st)[a]	4.4%[†ALL]	19.6%[†1]	17.8%[†1]
% failed a consistent response to repeated choice task (2nd)[a]	2.5%[†3]	46.8%[b]	19.1%[†1]
% who did not provide same answer for best and worst (1st)		58.9%	61.2%
% who did not provide same answer for best and worst (2nd)		72.8%[§]	60.5%
Focus on top and bottom levels			
Mean individual % of choosing level 1 as best		70.5%[†3]	59.9%[†2]
Mean individual % of choosing level 5 as worst		76.3%	69.4%

Note: [†] $p < 0.10$ of chi-squared tests comparing study arms 1, 2, and 3 (if applicable).
[a] For BWS defined as providing either the same best or worst answer.
[b] Choice task with intensity colour coding being reversed.

number of choice tasks participants could handle.

3.3. Choice strategies and choice behaviour

Most respondents strongly agreed with the statement that they compared all dimensions/items before making their choices, with no significant differences between study arms (Table 2). There were mixed results concerning the use of simplifying choice heuristics or strategies comparing DCE and BWS study arms. While DCE participants agreed to a lesser extent that they decided that all dimensions/items are equally important, they also reported to a larger degree to having based their decisions on the same 1 or 2 well-being dimensions, which implies some level of attribute non-attendance.

Table 3 lists results for the analysis of choice behaviour. The lexicographic score (see section 2.5), was significantly lower in DCE respondents, indicating more trading and less dominant choice behaviour. In the DCE, dominant preferences were observed only for the physical health attribute. In the BWS, such behaviour was also observed for the mental health and making ends meet attributes, with physical health still being the most prevalent one.

In the DCE study arm, 4.4% of respondents did not provide the same answer to the repeated choice task, when it appeared again for the first time (position 2 and 8), with the same colour code. When it was repeated again as the last choice task, that share was 2.5%. Up to 20% of respondents did not provide either the same best or worst answer in the repeated BWS tasks.[7] When defining consistency as providing the same answer to both best and worst, this share increased to around 60%. There were no significant differences between BWS study arms regarding the choice consistency of the first repeated instance. Almost half of respondents did not provide a consistent best or worst answer to the repeated BWS choice task, where the intensity colour coding was reversed (position 13). This share was 72.8% when defining consistency in terms of selecting the same best and worst items.

We further calculated the percentage of best and worst answers based on either the top and bottom levels of the WOOP dimensions on individual level and aggregated that by taking the average. The average share was between 60 and 75%, with higher values observed for the colour coded BWS tasks (significant difference for 'best').

4. Discussion

To assess the cognitive burden of different types of choice tasks for valuing well-being states for quality of life measures in older people, a randomised experiment was conducted, allocating respondents to either a DCE, a colour coded BWS, or a regular BWS format using an online setting. Our study contributes to the literature by providing empirical evidence on 1) whether DCE or BWS choice tasks are associated with lower cognitive burden in the context of health or well-being state valuation in an older population sample, and 2) whether colour coding of BWS tasks affects cognitive burden and to a lesser extent validity of BWS experiments.

Finding a lower drop-out rate and lower choice task completion time in the DCE study arm compared to the BWS study arms implies that, for older people, DCE choice tasks are less tiring and faster to complete than BWS tasks. Lower completion time was also observed by van Dijk et al. (2016). In terms of self-reported measures, our results indicate that the DCE tasks also were perceived as less cognitively burdensome, and that a higher number of DCE choice tasks was regarded as more acceptable than was a higher number of BWS tasks. The former has also been reported in related studies in different contexts (Whitty and Oliveira Gonçalves, 2018). The latter is especially relevant to consider when thinking about the number of choices per respondent, and hence the required sample size, when selecting DCE or BWS format. Finding lower cognitive burden associated with DCE tasks compared to BWS tasks, in general, is at odds with what has been reported before (Netten et al., 2012). The authors of that study also compared cognitive burden of DCE and BWS tasks for valuing a large descriptive system of a quality of life instrument, but the design of their study was fairly different. The authors used cognitive interviewing, a qualitative approach, in a small sample (N = 30), split the DCE task into two parts to reduce the

[7] It has to be acknowledged, though that the likelihood of providing the same answer by chance alone is larger for DCE choice tasks (50%).

difficulty of the task and showed both DCE and BWS tasks to respondents.[8] Whether the difference in findings relates to the differences in design of the studies, is difficult to say.

In terms of (simplifying) choice strategies and choice behaviour, which co-occur with larger cognitive burden, our results are mixed regarding the self-reported behaviour, and less clear cut. We did observe a considerably higher choice consistency and lower degrees of dominant choice behaviour for DCE respondents, with their measurement to some degree accommodating for the methodological differences. However, these results may relate more to artefacts of the type of choice task and may be unrelated to cognitive burden. As stated also by Whitty and Oliveira Gonçalves (2018), the probability of answering consistently to a DCE task by pure chance is already 50%. With nine dimensions this probability is much lower (22%) for the BWS task (defined as providing either the same best or worst answer). Nevertheless, finding that around 60% of BWS respondents did not provide the same best and worst answers when a choice was repeated for the first and the second time, is somewhat worrisome on its own. A higher degree of trading and lower degrees of dominant choice behaviour in DCEs were also reported in the related literature before (Krucien et al., 2017; Whitty et al., 2014) with a similar caveat as for analysing choice consistency.

Comparing colour coded with non-colour coded BWS, we found a similar drop-out rate for both tasks (1.9% and 3.2%, respectively). In the study by Jonker et al. (2018) (study arms 1 and 2), colour coding of the DCE tasks decreased the dropout rate from 13.9% to 9.8%. Further results from the same study set up showed that colour coding alone did not lead to differences with respect to the self-reported cognitive debriefing questions (Jonker et al., 2019). Our results for BWS regarding these questions are mixed. While participants of the colour coded BWS on average agreed to a higher extent that they could have answered more choice tasks, the non-colour coded BWS choice tasks appeared to have been clearer to respondents. Given no conclusive evidence on cognitive burden, and the fact that the colour coding increased the already high focus on top and bottom levels of the quality of life instrument in the BWS tasks, colour coding BWS cannot be recommended for health or well-being state valuation studies among older people.

The overall implications of our analysis must be interpreted considering several limitations. First, the rather small sample size did not provide us with enough statistical power to be able to use several blocks of choice tasks, which then also would have allowed us to estimate DCE and BWS models. During the design stage, we aimed for 150 respondents per study arm due to the small overall pool of individuals aged 65 on online platforms. While the choice sets were created according to standard design methodology, it could be the case that either of the two choice sets is more difficult to answer in general, irrespective of choice task format, due to smaller utility difference within the shown profiles. As utility weights for the WOOP are not available yet, it was not possible to account for that in the selection of choice set. This risk could have been reduced if multiple blocks would have been used. A second, related, limitation is that DCE and BWS models could not be estimated, which prevented us from analysing the actual choices people made. Testing for choice consistency or overall noise in the data would have given us an indication on the quality of the responses However, such a comparison between DCE and BWS responses would have come with additional limitations.

In terms of the generalisability of our results, we need to acknowledge the following: Our study was conducted in an online setting, with respondents from an online panel. As certain subpopulations with varying levels of cognitive abilities may self-select into such panels (especially in older ages), the representativity to the general population aged 65 and above may be limited. However, the purpose of our study was to provide an indication of cognitive burden of different methods *specifically* using respondents from online panels, which by now are the most frequently used sampling formats for these types of analyses (Mulhern et al., 2018). Therefore, our results should only be generalised to similar online settings. Our sample likely was on the upper end of the spectrum of cognitive abilities of people aged 65 and above (highly educated and rather healthy, see Table 1). It is not certain, whether our conclusions would be the same in a sample with average or low levels of cognitive abilities, as we did not measure cognitive abilities directly. However, using years of education as an imperfect proxy for overall cognitive abilities, we could not observe an education, and therefore cognitive ability, gradient in our results (i.e. the direction of our results remained stable, when splitting our sample into a lower and a higher educated group). To increase the representativeness of the sample in a full-scale valuation study among the elderly using online panels, it will be necessary to implement further age stratification by setting appropriate age group quotas.

As for the generalisability towards other online panels, the following limitation applies: Per online platform rule, the recruitment of respondents involved a monetary compensation which is rather high compared to standard online panels, and which can be reduced if the researcher is not satisfied with the quality of responses. While this is a good thing for respondents and their motivation, this led to very low dropout rates and could have also affected other parts of the analysis Another caveat of our analysis is that the applicability of our results to the comparison of DCEs without overlap and colour coding, and BWS is limited. However, the use of level overlap in similar DCEs as strategy to reduce task complexity seems to be increasing (e.g. King et al., 2018; Mulhern et al., 2019). Not really a limitation, but important to note in terms of cognitive burden is the following: In the DCE setup, it was possible and logical to reduce the level descriptions compared to the full level text in the BWS, as the attributes were already included on the left side of the task (Fig. 3). This may also have contributed to DCE tasks being perceived to be easier to handle.

5. Conclusions

Overall, we found evidence that level overlapped, and colour coded DCE choice tasks are less cognitively burdensome than BWS choice tasks, in a complex health (or, here, well-being) state valuation exercise among older people in an online setting. This has implications for future valuation studies, especially since the complexity of the measures to be valued seems to increase when moving

[8] Although it does not become clear from the paper, whether respondents had to answer full sets of choice tasks or only one task per method.

from health-related to overall quality of life; see, for instance, the WOOP (Appendix A), the current plans of the E-QALY project (https://scharr.dept.shef.ac.uk/e-qaly/), or another ongoing study developing a quality of life measure for older people (Ratcliffe et al., 2019). Cognitive burden should be an important factor in deciding about which method to choose for valuing such descriptive systems, but at the same time, statistical and theoretical aspects need to be considered as well. Although our results may not be easily generalisable to other topics of study within or outside health care and to other study populations, our analysis may at least serve as a good example of how to assess cognitive burden associated with different types of choice experiments.

Funding

Sebastian Himmler receives financial support from a grant from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 721402). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Author contributions

Sebastian Himmler: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **Vikas Soekhai:** Validation, Writing - Original Draft, Writing - Review & Editing, Visualization **Job van Exel:** Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision **Werner Brouwer:** Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank members of the Erasmus Choice Modelling Center (ECMC) for valuable feedback during two seminars in which this work was presented. We would also like to thank Marcel Jonker, who provided us with a template for programming the survey.

Appendix

Appendix A

Well-being of Older People (WOOP) instrument

For each section, select the description that is most appropriate for you today.

Physical health. Consider physical conditions or ailments and other physical impairments that affect your daily functioning.

- I have no problems with my physical health
- I have slight problems with my physical health
- I have moderate problems with my physical health
- I have severe problems with my physical health
- I have very severe problems with my physical health

Mental health. Consider problems with your ability to think, anxiety, depression and other mental impairments that affect your daily functioning.

- I have no problems with my mental health
- I have slight problems with my mental health
- I have moderate problems with my mental health
- I have severe problems with my mental health
- I have very severe problems with my mental health

Social life. Consider your relationship with your partner, family or other people who are important to you. This concerns the amount and quality of the contact you have.

- I'm very satisfied with my social life
- I'm satisfied with my social life
- I'm reasonably satisfied with my social life
- I'm dissatisfied with my social life
- I'm very dissatisfied with my social life

Receive support. Everyone needs help or support sometimes. Consider practical or emotional support, for example from your partner, family, friends, neighbours, volunteers or professionals. This concerns being able to count on support when you need it, as well as the quality of the support.

- I'm very satisfied with the support I get, when needed
- I'm satisfied with the support I get, when needed
- I'm reasonably satisfied with the support I get, when needed
- I'm dissatisfied with the support I get, when needed
- I'm very dissatisfied with the support I get, when needed

Acceptance and resilience. Consider your acceptance of your current circumstances and your ability to adapt to changes to these, whether or not with support of your religion or belief.

- I'm very able to deal with my circumstances and changes to these
- I'm able to deal with my circumstances and changes to these
- I'm reasonably able to deal with my circumstances and changes to these
- I'm not able to deal with my circumstances and changes to these
- I'm not at all able to deal with my circumstances and changes to these

Feeling useful. Consider meaning something to others, your environment or a good cause.

- I feel very useful
- I feel useful
- I feel reasonably useful
- I do not feel useful
- I do not feel at all useful

Independence. Consider being able to make your own choices or doing the activities that you find important.

- I feel very independent
- I feel independent
- I feel reasonably independent
- I feel dependent
- I feel very dependent

Making ends meet. Consider having enough money to meet your daily needs and having no money worries.

- I'm more than able to make ends meet
- I'm able to make ends meet
- I'm reasonably able to make ends meet
- I'm not able to make ends meet
- I'm not at all able to make ends meet

Living situation. Consider living in a house or neighbourhood you like.

- I'm very satisfied with my living arrangements
- I'm satisfied with my living arrangements
- I'm reasonably satisfied with my living arrangements
- I'm dissatisfied with my living arrangements
- I'm very dissatisfied with my living arrangements

Appendix B

Repeated choice tasks

Which of the described well-being states do you prefer, A or B? *(2 of 6)*

	A	B
Physical health	**Moderate** problems	**Moderate** problems
Mental health	**Moderate** problems	**Severe** problems
Social contacts	**Very dissatisfied**	**Reasonably satisfied**
Receiving support	**Very satisfied**	**Very dissatisfied**
Acceptance	**Almost unable** to cope	**Almost unable** to cope
Feeling useful	Feeling **very** useful	Feeling **very** useful
Independency	Feeling **very** independent	Feeling **very** independent
Making ends meet	**Well** able to meet ends	**Well** able to meet ends
Living situation	**Satisfied**	**Very dissatisfied**
	◯	◯

- Positive aspects are **light blue** and negative aspects are **darker purple**
- Put the cursor above the underlined items for descriptions

Imagine living in this well-being state and select which aspect you would most prefer, and which aspect you would least prefer. *(2 of 6)*

Most	Well-being state	Least
◯	I am satisfied with my social contacts	◯
◯	I am dissatisfied with the support I receive	◯
◯	I am almost unable to deal with my circumstances and changes therein	◯
◯	I feel useful	◯
◯	I have very severe problems with my physical health	◯
◯	I have no problems with my mental health	◯
◯	I feel independent	◯
◯	I am very well able to make ends meet	◯
◯	I am satisfied with my living situation	◯

- Positive aspects are **light blue** and negative aspects are **darker purple**
- Put the cursor above the underlined items for descriptions

Appendix C

Utility function for DCE design

The following utility function was optimised in Ngene, where i indicates the respondent and j the well-being profile:

$$U_{ij} = PH_{ij}\beta_{PH} + MH_{ij}\beta_{MH} + SOC_{ij}\beta_{SOC} + SUP_{ij}\beta_{SUP} + ACC_{ij}\beta_{ACC} + USE_{ij}\beta_{USE} + IND_{ij}\beta_{IND} + MEM_{ij}\beta_{MEM} + LIV_{ij}\beta_{LIV} + \varepsilon_{ij} \quad (1)$$

PH, MH, SOC, SUP, ACC, USE, IND, MEM, and LIV symbolise vectors of the levels of the WOOP instrument (Appendix A). The betas represent vectors of four parameters each, which model the utility associated with each of the levels of the nine dimensions of the WOOP compared to the lowest level in each dimension.

References

Arendts, G., Jan, S., Beck, M.J., Howard, K., 2017. Preferences for the emergency department or alternatives for older people in aged care: a discrete choice experiment. Age Ageing 46, 124–129. https://doi.org/10.1093/ageing/afw163.

Cheung, K.L., Wijnen, B.F.M., Hollin, I.L., Janssen, E.M., Bridges, J.F., Evers, S.M.A.A., Hiligsmann, M., 2016. Using best–worst scaling to investigate preferences in health care. Pharmacoeconomics 34, 1195–1209. https://doi.org/10.1007/s40273-016-0429-5.

Coast, J., Flynn, T.N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J.J., Peters, T.J., 2008. Valuing the ICECAP capability index for older people. Soc. Sci. Med. 67, 874–882. https://doi.org/10.1016/j.socscimed.2008.05.015.

de Meijer, C., Wouterse, B., Polder, J., Koopmanschap, M., 2013. The effect of population aging on health expenditure growth: a critical review. Eur. J. Ageing 10, 353–361. https://doi.org/10.1007/s10433-013-0280-x.

de Winter, J., Dodou, D., 2010. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012). Practical Assess. Res. Eval. 15 https://doi.org/10.7275/bj1p-ts64.

Devlin, N.J., Shah, K.K., Feng, Y., Mulhern, B., van Hout, B., 2018. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 27, 7–22. https://doi.org/10.1002/hec.3564.

Flynn, T.N., 2010. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Rev. Pharmacoecon. Outcomes Res. 10, 259–267. https://doi.org/10.1586/erp.10.29.

Flynn, T.N., Huynh, E., Peters, T.J., Al-Janabi, H., Clemens, S., Moody, A., Coast, J., 2015. Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. Health Econ. (United Kingdom) 24, 258–269. https://doi.org/10.1002/hec.3014.

Flynn, T.N., Marley, A.A.J., 2014. Best-worst scaling: theory and methods. In: Handbook of Choice Modelling. Edward Elgar Publishing, pp. 178–201. https://doi.org/10.4337/9781781003152.00014.

Franco, M.R., Howard, K., Sherrington, C., Rose, J., Ferreira, P.H., Ferreira, M.L., 2016. Smallest worthwhile effect of exercise programs to prevent falls among older people: estimates from benefit-harm trade-off and discrete choice methods. Age Ageing 45, 806–812. https://doi.org/10.1093/ageing/afw110.

Hackert, M.Q.N., Brouwer, W.B.F., Hoefman, R.J., van Exel, J., 2019. Views of older people in The Netherlands on wellbeing: a Q-methodology study. Soc. Sci. Med. 240, 112535. https://doi.org/10.1016/j.socscimed.2019.112535.

Huynh, E., Coast, J., Rose, J., Kinghorn, P., Flynn, T., 2017. Values for the ICECAP-Supportive Care Measure (ICECAP-SCM) for use in economic evaluation at end of life. Social Science & Medicine 189, 114–128. https://doi.org/10.1016/j.socscimed.2017.07.012.

Jonker, M.F., Attema, A.E., Donkers, B., Stolk, E.A., Versteegh, M.M., 2017. Are health state valuations from the general public biased? A test of health state reference dependency using self-assessed health and an efficient discrete choice experiment. Health Econ. (United Kingdom) 26, 1534–1547. https://doi.org/10.1002/hec.3445.

Jonker, M.F., Donkers, B., de Bekker-Grob, E., Stolk, E.A., 2019. Attribute level overlap (and colour coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments. Health Econ. (United Kingdom) 28, 350–363. https://doi.org/10.1002/hec.3846.

Jonker, M.F., Donkers, B., de Bekker-Grob, E.W., Stolk, E.A., 2018. Effect of level overlap and colour coding on attribute non-attendance in discrete choice experiments. Value Health 21, 767–771. https://doi.org/10.1016/j.jval.2017.10.002.

King, M.T., Viney, Rosalie, Simon Pickard, A., Rowen, Donna, Aaronson, N.K., Brazier, J.E., Cella, David, Costa, D.S.J., Fayers, P.M., Kemmler, Georg, McTaggart-Cowen, H., Mercieca-Bebber, Rebecca, Peacock, S., Street, D.J., Young, T.A., Norman, Richard, Aaronson, N., Brazier, J., Cella, D., Costa, D., Fayers, P., Grimison, P., Janda, M., Kemmler, G., King, M., McTaggart-Cowan, H., Mercieca-Bebber, R., Norman, R., Pickard, S., Rowen, D., Velikova, G., Viney, R., Street, D., Young, T., 2018. Australian utility weights for the EORTC QLU-C10D, a multi-attribute utility instrument derived from the cancer-specific quality of life questionnaire, EORTC QLQ-C30. Pharmacoeconomics 36, 225–238. https://doi.org/10.1007/s40273-017-0582-5.

Krucien, N., Watson, V., Ryan, M., 2017. Is best–worst scaling suitable for health state valuation? A comparison with discrete choice experiments. Health Econ. (United Kingdom) 26, e1–e16. https://doi.org/10.1002/hec.3459.

Louviere, J.L., 2004. Random utility theory-based stated preference elicitation methods: applications in health economics with special reference to combining sources of preference data. In: Working Paper at Center for the Study of Choice No. 04-001.

Maddala, T., Phillips, K.A., Johnson, F.R., 2003. An experiment on simplifying conjoint analysis designs for measuring preferences. Health Econ. 12, 1035–1047. https://doi.org/10.1002/hec.798.

Makai, P., Brouwer, W.B.F., Koopmanschap, M.A., Stolk, E.A., Nieboer, A.P., 2014. Quality of life instruments for economic evaluations in health and social care for older people: a systematic review. Soc. Sci. Med. 102, 83–93. https://doi.org/10.1016/j.socscimed.2013.11.050.

Milte, R., Ratcliffe, J., Chen, G., Lancsar, E., Miller, M., Crotty, M., 2014. Cognitive overload? An exploration of the potential impact of cognitive functioning in discrete choice experiments with older people in health care. Value Health 17, 655–659. https://doi.org/10.1016/j.jval.2014.05.005.

Mulhern, B., Norman, R., De Abreu Lourenco, R., Malley, J., Street, D., Viney, R., 2019. Investigating the relative value of health and social care related quality of life using a discrete choice experiment. Soc. Sci. Med. 233, 28–37. https://doi.org/10.1016/j.socscimed.2019.05.032.

Mulhern, B., Norman, R., Street, D.J., Viney, R., 2018. One method, many methodological choices: a structured review of discrete-choice experiments for health state valuation. Pharmacoeconomics. https://doi.org/10.1007/s40273-018-0714-6.

Netten, A., Burge, P., Malley, J., Potoglou, D., Towers, A-M, Brazier, J., Flynn, T., Forder, J., Wall, B., 2012. Outcomes of social care for adults: developing a preference-weighted measure. Health Technology Assessment 16 (16). https://doi.org/10.3310/hta16160.

Neumann, P.J., Sanders, G.D., Russell, L.B., Siegel, J.E., Ganiats, T.G., 2016. Cost Effectiveness in Health and Medicine. Oxford University Press, New York.

Palan, S., Schitter, C., 2018. Prolific.ac—a subject pool for online experiments. J. Behav. Exp. Financ. 17, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004.

Potoglou, D., Burge, P., Flynn, T., Netten, A., Malley, J., Forder, J., Brazier, J.E., 2011. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. Soc. Sci. Med. 72, 1717–1727. https://doi.org/10.1016/j.socscimed.2011.03.027.

Ratcliffe, J., Cameron, I., Lancsar, E., Walker, R., Milte, R., Hutchinson, C.L., Swaffer, K., Parker, S., 2019. Developing a new quality of life instrument with older people for economic evaluation in aged care: study protocol. BMJ Open 9, e028647. https://doi.org/10.1136/bmjopen-2018-028647.

Ryan, M., Gerard, K., Amaya-Amaya, M. (Eds.), 2008. Using Discrete Choice Experiments to Value Health and Health Care, the Economics of Non-market Goods and Resources. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-1-4020-5753-3.

Soekhai, V., de Bekker-Grob, E.W., Ellis, A., Vass, C.M., 2019. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics 37, 201–226. https://doi.org/10.1007/s40273-018-0734-2.

van Dijk, J.D., Groothuis-Oudshoorn, C.G.M., Marshall, D.A., IJzerman, M.J., 2016. An empirical comparison of discrete choice experiment and best-worst scaling to estimate stakeholders' risk tolerance for Hip Replacement surgery. Value Health 19, 316–322. https://doi.org/10.1016/j.jval.2015.12.020.

Whitty, J.A., Oliveira Gonçalves, A.S., 2018. A systematic review comparing the acceptability, validity and concordance of discrete choice experiments and best–worst scaling for eliciting preferences in healthcare. Patient 11, 301–317. https://doi.org/10.1007/s40271-017-0288-y.

Whitty, J.A., Walker, R., Golenko, X., Ratcliffe, J., 2014. A think aloud study comparing the validity and acceptability of discrete choice and best worst scaling methods. PloS One 9. https://doi.org/10.1371/journal.pone.0090635.

Yao, R.T., Scarpa, R., Rose, J.M., Turner, J.A., 2015. Experimental design criteria and their behavioural efficiency: an evaluation in the field. Environ. Resour. Econ. 62, 433–455. https://doi.org/10.1007/s10640-014-9823-7.