UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

BUSINESS SCHOOL
Te Kura Pakihi

ISSN 1178-2293 (Online)

University of Otago
Economics Discussion Papers
No. 1903

FEBRUARY 2019

A new tool for creating personal and social EQ-5D-5L value sets, including valuing 'dead'

Trudy Sullivan,[1*] Paul Hansen,[2,3] Franz Ombler,[3] Sarah Derrett,[1] Nancy Devlin[4]

[1]Department of Preventive and Social Medicine, University of Otago, Dunedin, New Zealand [2]Department of Economics, University of Otago, Dunedin, New Zealand

[3]1000minds Ltd, Wellington, New Zealand

[4]Centre for Health Policy, University of Melbourne, Melbourne, Australia.

Address for correspondence:

Paul Hansen
Department of Economics
University of Otago
PO Box 56
Dunedin
NEW ZEALAND
Email: paul.hansen@otago.ac.nz
Telephone: 64 3 479 8547

A new tool for creating personal and social EQ-5D-5L value sets, including valuing ‘dead’

Trudy Sullivan,[1*] Paul Hansen,[2,3] Franz Ombler,[3] Sarah Derrett,[1] Nancy Devlin[4]

[1]Department of Preventive and Social Medicine, University of Otago, Dunedin, New Zealand

[2]Department of Economics, University of Otago, Dunedin, New Zealand

[3]1000minds Ltd, Wellington, New Zealand

[4]Centre for Health Policy, University of Melbourne, Melbourne, Australia.

[*]*Correspondence*: Trudy Sullivan, Department of Preventive and Social Medicine, University of Otago, Dunedin 9054, New Zealand. E-mail: trudy.sullivan@otago.ac.nz; Ph: +64 3 479 8087.

Abstract

A new online tool for creating *personal* and *social* EQ-5D-5L value sets was recently developed and trialled in New Zealand (NZ). Health state values for each participant are determined using the PAPRIKA method – in the present context, a novel type of adaptive discrete choice experiment – and any health states worse than dead are identified using a binary search algorithm. Following testing and refinement, the tool was distributed in an online survey to a representative sample of NZ adults (N=5112), whose personal value sets were created. Extensive data quality checks were performed, resulting in a ‘high-quality’ sub-sample of 2468 participants whose personal value sets were, in effect, averaged to create a social value set for NZ, as represented by social ‘disutility coefficients’ (consistent with the EQ-5D literature). These results overall and participants’ feedback indicate that the new valuation tool is feasible and acceptable to participants and enables valuation data to be relatively easily and cheaply collected. The tool could also be used in other countries, tested against other methods for creating EQ-5D-5L value sets, applied in personalised medicine and adapted to create value sets for other health descriptive systems.

Key words: EQ-5D-5L, value sets, health-related quality of life (HRQoL), health state preferences, discrete choice experiment (DCE), PAPRIKA method, binary search, bisection, 1000minds, New Zealand

Acknowledgments: Thank you to Josh Ward for helping to develop the survey and facilitating the think-aloud sessions and snowball sample, to our colleagues, friends and families who took part in the pilot-testing, participants in the think-aloud sessions and everyone who completed the snowball and general population surveys. Thanks also for feedback received at the 35th EuroQol Group Scientific Plenary Meeting 2018. The support of the Department of Preventive and Social Medicine, by providing funding for data collection and software administration, and 1000minds Ltd, for supplying the software used for the valuation tool, is also gratefully acknowledged.

Conflicts of interest: The second and third authors have ownership interests in 1000minds Ltd; the fourth and fifth authors are members of the EuroQol Group.

1. INTRODUCTION

The EuroQol Group's health descriptive systems, the EQ-5D-3L (Brooks, 1996) and EQ-5D-5L (Herdman, et al., 2011), are the most widely used instruments worldwide (Richardson, McKie, & Bariola, 2017) for representing health-related quality of life (HRQoL) for use in cost-utility analysis (CUA) and patient-reported health outcome measures (PROMs). Introduced in 2009, the EQ-5D-5L is progressively superseding the original EQ-5D-3L. Both versions have five HRQoL dimensions – mobility, self-care, usual activities, pain/discomfort and anxiety/depression – but instead of just three levels of severity for the EQ-5D-3L, the EQ-5D-5L has five levels: e.g. no, slight, moderate, severe and extreme problems. Thus, the EQ-5D-5L represents 3125 (5^5) health states, compared to just 243 (3^5) for the EQ-5D-3L. This increased granularity comes at the expense of higher complexity and cost to create EQ-5D-5L value sets relative to its simpler predecessor.

An increasingly common approach for creating patient-reported outcome instruments, including EQ-5D-5L value sets, is to use discrete choice experiments (DCEs) (McFadden, 1973). In the present context, DCEs involve participants expressing their preferences with respect to the relative importance of EQ-5D-5L dimensions by repeatedly choosing between hypothetical health states. Compared to other choice-based valuation techniques (e.g. time trade-off and standard gamble methods), DCEs are cognitively less challenging – typically requiring participants to choose between just two health states at a time – and so DCEs can be implemented relatively easily and cheaply using online surveys (e.g. Krabbe et al., 2014).

DCEs have been used to create EQ-5D-5L value sets for: Australia (Norman, Cronin, & Viney, 2013; Viney et al., 2014), Japan (Shiroiwa et al., 2016), South Korea (Kim et al., 2016), Netherlands (Versteegh et al., 2016), Spain (Ramos-Goñi et al., 2017), Germany (Ludwig, von der Schulenburg, & Greiner, 2018), United Kingdom (Mulhern, Bansback, Hole, & Tsuchiya, 2017) and England (Devlin, Shah, Feng, Mulbern, & van Hout, 2018; Feng, Devlin, Shah, Mulbern, & van Hout, 2018).

Most of these value sets were created using the EuroQol Group's software-supported protocol, known as the "EuroQoL Valuation Technology" (EQ-VT) (Oppe, Devlin, van Hout, Krabbe, & de Charro, 2014). The EQ-VT is interviewer-administered and comprises 10 time trade-off (TTO) questions, including a lead-time TTO for states worse than dead (Robinson, & Spencer, 2006; Devlin et al., 2013), and seven DCE questions involving choosing between two hypothetical health states defined on all five dimensions at a time (i.e. a 'full profile' DCE).

The EQ-VT protocol and the quality of the resulting valuation data have been challenged. Stolk, Ludwig, Rand, van Hout, & Ramos-Goñi (2019) discusses problems with the original version of the EQ-VT used in the first wave of valuation studies (in Canada, England, China and the Netherlands) and subsequent mitigations in later versions, including tighter quality control over interviewers. Future modifications are mooted, including less reliance on the TTO method and more on DCEs. A formal review of the EQ-5D-5L value set for England, commissioned by the Department of Health for England, criticises the EQ-VT (including its latest version) and the quality of the data and raises serious concerns about the TTO method (Hernández-Alva, Pudney, & Wailoo, 2018). "[T]he large number of [problematic TTO responses] and the egregious nature of some of the anomalies suggest that there might be serious difficulties for participants relating to their engagement with or understanding of the TTO tasks." (p. 35). The report recommends: "A new programme of further development, including a new data collection initiative [i.e. beyond the EQ-VT protocol], should be considered to put EQ-5D-5L on a sufficiently firm evidential basis." (p. 58).

This paper presents a new DCE-based online tool for creating EQ-5D-5L value sets that includes extensive checks of the quality of participants' data. The tool implements a novel type of adaptive DCE in the HRQoL context, known as the PAPRIKA method (Hansen & Ombler, 2008). A major advantage of the PAPRIKA method relative to most other DCE methods is that, as well as producing a *social* value set, PAPRIKA immediately (and in 'real time') produces a *personal* value set for each participant. This capability represents a unique opportunity to personalise value sets, consistent with personalised (precision) medicine (Mirnezami, Nicholson, & Darzi, 2012) and the increasing interest in patient preferences. Another major advantage of individual-level data is the opportunity for examining the heterogeneity of individual health state preferences and differences between population sub-groups – e.g. healthy vs chronically ill, young vs old, ethnic group differences, etc. A third advantage of individual value sets is that a binary search algorithm can be used for participants to identify any health states worse than dead (e.g. Devlin, Shah, Mulhern, Pantiri, & van Hout, 2018), as in the tool presented here.

The objective of this paper is to present the new valuation tool, especially the DCE and binary search algorithm, and to report on the tool's feasibility and acceptability to participants. The setting for developing and trialling the tool is New Zealand (NZ). As well as the authors being from there, the EQ-5D-5L's adoption internationally suggests the need for a NZ value set given the EQ-5D-3L value set created in 1999 (Devlin, Hansen, Kind & Williams, 2003) is used extensively by researchers and policy-makers – e.g. for CUA by NZ's Pharmaceutical Management Agency (PHARMAC, 2017) and, to a lesser extent, for PROMs by the Ministry of Health (Shuker et al., 2017). Hence, results for creating a NZ social value set are also presented.

2. METHODS

2.1 Tool development and testing

The new valuation tool has two main components: a DCE to determine each participant's 3125 health state values, and a binary search algorithm to identify health states worse than dead. After informal testing, a 'think-aloud' approach (Gilhooly & Green, 1996) was used to assess the feasibility and acceptability of the DCE and two methods for identifying health states worse than dead: a binary search algorithm (explained in sub-section 2.3) and an interactive visual analogue scale (VAS). An example of a VAS question appears in Appendix 1, along with an earlier presentation of the binary search question (superseded by the final version below).

Twelve one-hour think-aloud sessions were held. In the presence of two interviewers, including two of the authors (TS or SD), the participant verbalised their thoughts while using the tool, with the interviewer asking follow-up questions. In short (results to be reported elsewhere), participants found answering questions involving a binary choice between a health state and dead (Figure 2) much easier than using a VAS slider (Figure A1 in the Appendix) and had more confidence in their answers; therefore, the binary search algorithm was adopted for the tool. Other think-aloud feedback was used to improve the tool's instructions and user interface. The tool was pilot-tested on a snowball sample of 270 participants and further refined.

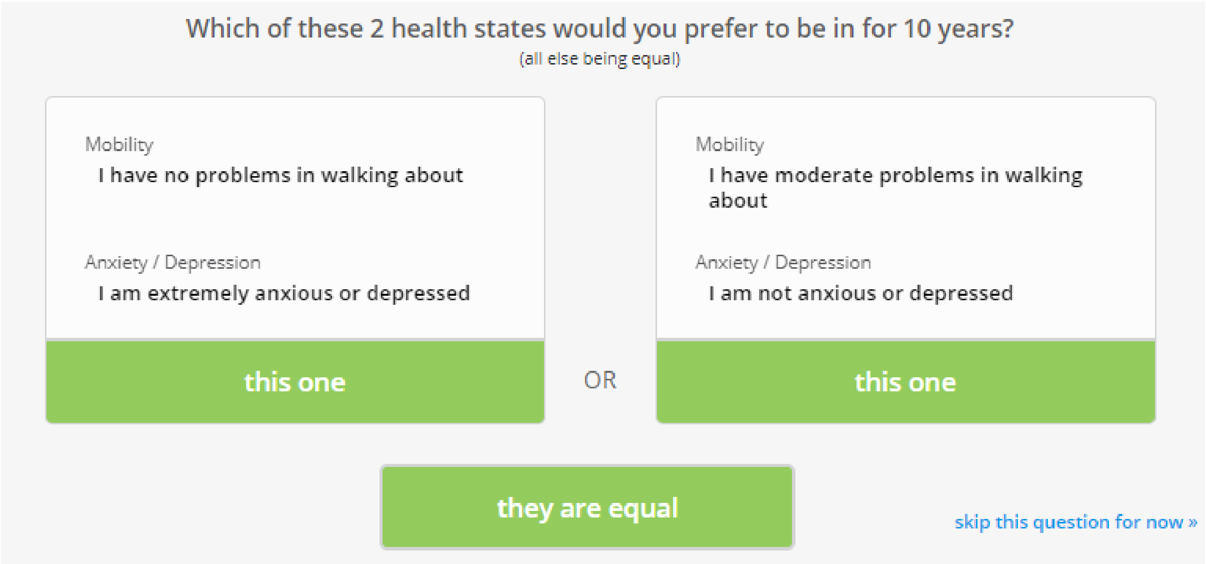
The tool's two main components are now explained in turn.

2.2 Discrete choice experiment

The DCE is based on the PAPRIKA method (Hansen & Ombler, 2008) – an acronym for ***P***otentially ***A***ll ***P***airwise ***R***an***K***ings of all possible ***A***lternatives – as implemented by 1000minds software (www.1000minds.com). Since 2004, this method and software have been used in a wide range of health decision-making areas, including health technology prioritisation (Martelli, et al., 2016; Sullivan & Hansen, 2017), patient prioritisation (Fitzgerald, et al., 2011; Hansen, Hendry, Naden, Ombler, & Stewart, 2012), disease classification and diagnosis (Shiboski et al., 2017), for developing clinical guidelines (French, et al., 2015) and prioritising diseases for R&D (Tacconelli et al., 2018).

The PAPRIKA method involves the participant repeatedly choosing between, in the present context, two hypothetical health states with respect to which one they would prefer to be in for 10 years. The health states are defined on two dimensions at a time (i.e. in effect, the other three dimensions are assumed to be the same) and involve a trade-off. An example of such a question appears in Figure 1.

FIGURE 1 Example of a DCE question from 1000minds software



Which of these 2 health states would you prefer to be in for 10 years?
(all else being equal)

Mobility
I have no problems in walking about

Anxiety / Depression
I am extremely anxious or depressed

Mobility
I have moderate problems in walking about

Anxiety / Depression
I am not anxious or depressed

Such questions are repeated with different pairs of hypothetical health states, always involving trade-offs between different combinations of the levels on the dimensions, two at a time. Each time a participant ranks a pair of states, all other states that can be pairwise ranked via transitivity are identified and eliminated, thereby minimising the number of questions asked. For example, if a person prefers state *A* to *B* and *B* to *C*, then, by transitivity, *A* is also preferred to *C* (so the person would not be asked to choose between *A* and *C*). Also, each time a person answers a question, based on that and all preceding answers, PAPRIKA adapts with respect to choosing their next question (always one whose answer is not implied by earlier answers). Based on the person's next answer, another question is presented; then another, and another; and so on. Thus, PAPRIKA is a type of *adaptive* DCE. Combined with the transitivity-based elimination procedure, this adaptivity minimises the number of questions a participant is asked while ensuring they end up having pairwise ranked all possible states defined on two dimensions at a time, either explicitly or implicitly (by transitivity).

Finally, from the participant's explicit pairwise rankings (i.e. answers to the DCE questions) the software uses linear programming techniques to derive weights for the levels on each dimension; for technical details, see Hansen & Ombler (2008). As well as each participant's weights, and their corresponding personal value set, these individual outputs are averaged across all participants to produce social weights and a social value set.

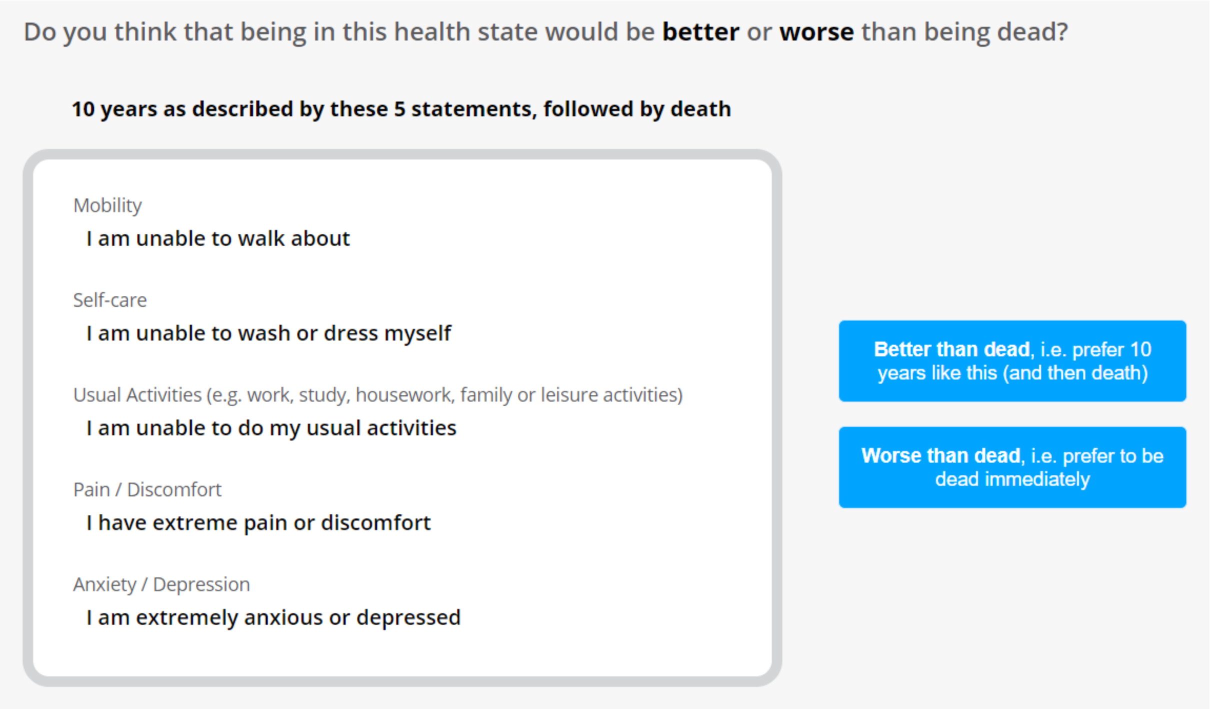
To restrict the number of DCE questions asked, only levels 1, 3 and 5 of each of the EQ-5D-5L dimensions were included in the DCE. The weights for levels 2 and 4 were interpolated using Bézier interpolation (Farin, Hoschek, & Kim, 2002) by, in essence, fitting a monotonic smoothed curve through the weights for levels 1, 3 and 5. Also, five combinations of levels (health states) deemed to be unrealistic to most people were suppressed: e.g. "no problems doing my usual activities" and either "extreme pain or discomfort" or "extremely anxious or depressed" or "unable to wash or dress myself" or "unable to walk about".

2.3 Binary search to identify health states worse than dead

An interactive binary search (or bisection) algorithm was implemented by the 1000minds software for participants to identify any health states they consider to be worse than dead. Such a search algorithm is possible because the PAPRIKA DCE method produces a *personal* value set for each participant.

The algorithm begins with the participant being asked if they think that being in the lowest-ranked heath state, 55555, for 10 years (the conventional time period) would be better or worse than being dead; this question is shown in Figure 2. If the person answers 55555 is "better than dead", the algorithm stops. If instead they answer 55555 is "worse than dead", the algorithm proceeds to search for, in effect, the 'dividing line' that splits their ranking of the 3125 states into ones better and worse than dead respectively.

FIGURE 2 Example of a binary search question to identify states worse than dead



Do you think that being in this health state would be **better** or **worse** than being dead?

10 years as described by these 5 statements, followed by death

Mobility
I am unable to walk about

Self-care
I am unable to wash or dress myself

Usual Activities (e.g. work, study, housework, family or leisure activities)
I am unable to do my usual activities

Pain / Discomfort
I have extreme pain or discomfort

Anxiety / Depression
I am extremely anxious or depressed

Better than dead, i.e. prefer 10 years like this (and then death)
Worse than dead, i.e. prefer to be dead immediately

Thus, if the person answers 55555 is “worse than dead”, they are asked if another, higher-ranked health state (set by the tool to 33333), is better or worse than dead. Depending on their answer, another higher- or lower-ranked state is evaluated: if the person answers 33333 is “worse than dead”, 22222 is posed next; if instead they answer 33333 is “better than dead”, 44444 is posed next. Having identified the range of health states in which dead lies, the algorithm proceeds to repeatedly bisect (halve) the participant’s personal ranking of states.

For example, suppose with reference to the questions above, the person answers 33333 is “better than dead” and then 44444 is “worse than dead”; they are then asked if the state in the middle of *their* ranking of 33333 to 44444 – e.g. perhaps 34432 – is better or worse than dead. Suppose they answer 34432 is “better than dead”; they are then asked if the state halfway in their ranking of 34432 to 44444 – e.g. perhaps 44433 – is better or worse than dead. This process continues: repeatedly halving the range of values until the dividing line is found that splits their ranking of the 3125 states into ones better and worse than dead.

In summary, three results with respect to the location of dead within the 3125 health states are possible: either dead is worse than 55555, and so dead and 55555 are both valued at 0 (customary for EQ-5D valuations); or 11111 is worse than dead (uncommon), and so dead = 1; or (most often) dead is spanned by two adjacent states in the person’s ranking (one better than, the other worse than dead), and so dead’s value (before rescaling) is the average of the values for these two adjacent states.

The cognitive effort associated with the binary search algorithm is considerably less than with more traditional implementations, in the following four ways.

One, the interface does not permit the health state under consideration to be valued as equal to dead (indifference); i.e. as illustrated in Figure 2, there is no “equal to dead” button. The think-aloud sessions and pilot-testing confirmed that presenting just two simple options reduces the cognitive difficulty and eliminates the risk that people might equate some states with dead even though, in fact, they think the states are (strictly) worse than dead.

Two, the algorithm begins by presenting 55555 rather than the state in the middle of a participant's personal value set. This feature means that the people for whom any minimal HRQoL (i.e. 55555) is better than dead need to answer only one question.

Three, for the first three questions asked by the algorithm, each participant is presented with 'balanced' states comprising identical levels (e.g. 33333) rather than the states exactly bisecting the participant's personal ranking of states. Because balanced states are easier for people to think about, they are less likely to make mistakes.

Four, instead of performing the binary search across all 3125 states, for simplicity a subset is used for each participant that still ensures reasonable accuracy. This subset (for each participant) is formed by grouping the 3125 states by their values rounded to two decimal places and selecting one from each group; thus, the subset has a maximum of 101 states with rounded values in the range 0-1. Consistent with 'balanced' states (as above), the state selected from each group has the lowest standard deviation across its levels (because this state is easier to think about); e.g. if 33233 and 12345 are in the same group, 33233 is selected.

Although participants' elicitation burden with respect to cognitive difficulty is reduced, the first three features above can increase the number of questions asked relative to a pure binary search algorithm; this number ranges from one (for a person answering 55555 is "better than dead") to a maximum of approximately nine.

2.4 New Zealand survey

International research company Research Now was engaged to recruit a sample representative of the NZ adult population with respect to age, gender, ethnicity and geographic location. Participants were asked to complete three parts of the online survey anonymously with respect to: (1) rating their current health status on the EQ-5D-5L questionnaire and EQ VAS; (2) completing the DCE and identifying health states worse than dead; and (3) supplying socio-demographic and background information and feedback about the tool and the survey overall. Ethics approval was obtained from the University of Otago Human Ethics Committee (D17/297).

2.5 Participant exclusions

The quality of participants' data is assessed in several ways. The objective is to identify people with 'low-quality' data and exclude them, thereby producing a sub-sample of participants with 'high-quality' data to use for creating the social value set.

First, participants whose solicited feedback indicates their survey responses are invalid were excluded. Also excluded were participants who value dead at unity – because rescaling their value set relative to 11111 = 1 and dead = 0 (see next sub-section) is mathematically impossible. In addition, the new tool includes the following data quality checks that can be used for identifying who else to exclude.

To test the consistency (reliability) of each participant's answers, at the end of the DCE the tool repeats two questions, selected according to the participant's strongest expression of their preferences so that they should be the easiest questions to answer. The software also records the time taken by each participant to answer their DCE questions. Thus, participants who fail to answer their two repeated questions identically (consistently) and/or who answer their questions implausibly quickly – i.e. below a threshold to be determined after scrutiny of the survey dataset – are excluded. The tool also records for each participant how many DCE questions are answered by clicking "this one" on the left-hand side versus "this one" on the right-hand side versus "they are equal" (see Figure 1 again). Finally, after being shown their ranking of the five dimensions produced by the DCE, participants are asked whether this ranking is as they expected.

2.6 Calculating personal and social value sets

Each participant's personal value set of 3125 health state values can be calculated by summing their DCE preference weights for each state's combination of levels on the five dimensions (where, by definition, states 11111 and 55555 are equal to unity and zero respectively). The value v of each health state can then be rescaled using the usual rescaling formula $\frac{v - z}{1 - z}$, where z is the participant's value for dead (from the binary search) (Oppe, Szende, & Charro, 2007).

Each participant's personal value set can also be represented by a linear equation analogous to the equations used to create social value sets in the EQ-5D literature (Oppe, Szende, & Charro, 2007). These equations comprise negative 'disutility' coefficients corresponding to level decrements from 'full health' state 11111 = 1, where health state values are calculated by adding unity to the sum of the negative coefficients corresponding to each state's combination of levels on the five dimensions. Each participant's disutility coefficients can be calculated from their DCE weights using the formula $\frac{w - w_1}{1 - z}$, where w is the weight for the level and dimension whose coefficient is being calculated, w_1 is the weight of level 1 (the highest weight) of the same dimension and z is the participant's value for dead.

From the participants' personal value sets, a social value set can be constructed by calculating the mean for each of the 3125 rescaled values across all individuals. The obvious method for calculating these mean values is to generate all participants' personal value sets and then calculate the mean value across all individuals for each health state. However, a simpler and equivalent method – i.e. less computationally intensive and resulting in the same social value set – is to average the above-mentioned personal disutility coefficients across all participants to determine the *social* disutility coefficients, from which the social value set is calculated by adding unity to the sum of the social coefficients for each state's combination of levels.

2.7 Correlations with other country EQ-5D-5L value sets

A cursory comparison of the NZ social value set with other countries' value sets is possible by calculating Spearman *ρ* and Pearson *r* correlation coefficients using other countries' value sets supplied by the EuroQol Group and applied in Ombler, Albert, & Hansen (2018). Significance levels discussed in the article[*] were used to test the coefficients' statistical significance.

3. RESULTS

The survey was completed by 5112 people (0.10% of NZ's population of 4.91 million). The full dataset is available on request from the authors.

Figure 3 represents the 2644 participant exclusions to obtain a high-quality sub-sample of 2468. Three participants whose feedback indicates their survey responses are invalid were immediately excluded. As represented by the Venn diagram in the figure, also excluded were participants who: value dead at unity, or are inconsistent (failed to answer the two repeated DCE questions identically), or are implausibly fast – i.e. a median time of less than six seconds per answer. This six second threshold (corresponding to participants answering half their DCE questions in under a minute) is supported by Figure 4, which reveals, overall, that the more time participants take to answer, the more consistent they are (also evident in Figure 3, where of the 2256 inconsistent participants, 683 are implausibly fast): at six seconds, most responses are consistent. In summary, of the 5112 who completed the survey, 2644 were excluded, leaving a sub-sample of 2468 for the analysis that follows.

In addition, a reassuring difference between the 2644 exclusions and the n=2468 sub-sample emerges with respect to people who answered all their DCE questions by clicking the same button. Of the 2644 exclusions, 225 (9%) always clicked the same button: mostly "they are equal". Given most of these 225 also answered inconsistently or implausibly quickly, it seems very likely that they carelessly 'clicked their way through' the DCE questions without due consideration. In contrast, just 132 (5%) in the n=2468 sub-sample always clicked the same button: "they are equal". A close examination of each of these people's data revealed nothing else 'suspicious' about their quality, including their answer to the question above about their DCE ranking of the five dimensions and any other general feedback, that might indicate they should also be excluded; and so, given always answering "they are equal" (i.e. indifference) is theoretically plausible, they were retained in the high-quality sub-sample.

[*] Available as an online resource from www.1000minds.com/sectors/health/hrqol.

FIGURE 3 Exclusions to obtain a high-quality sub-sample for creating a social value set (n=2468)

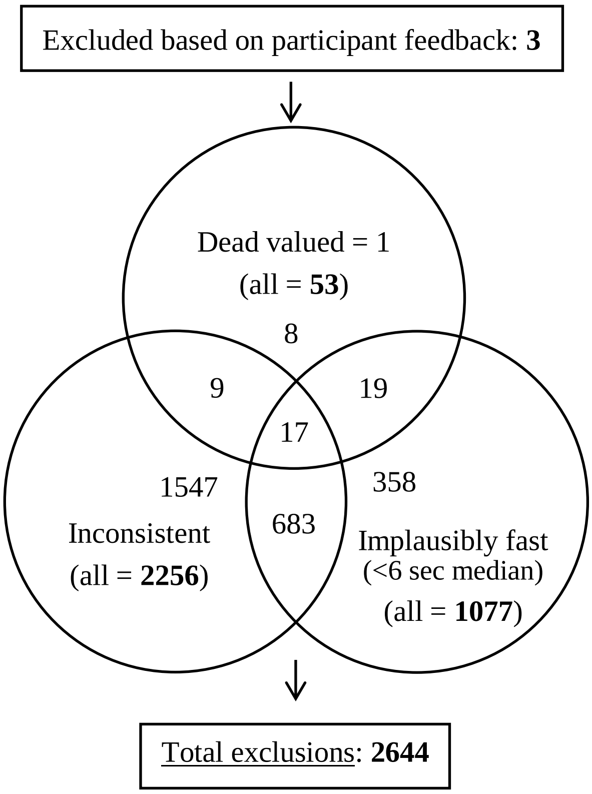
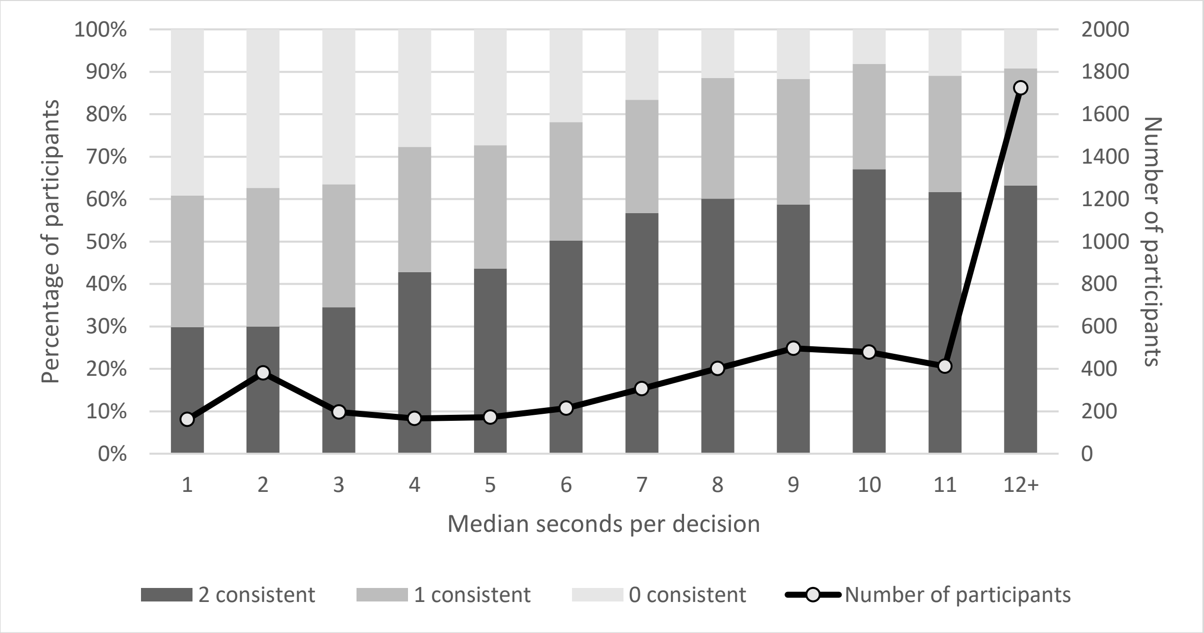


FIGURE 4 Median time per decision, compared with consistency of answers to two repeated DCE questions



The sub-sample's socio-demographic characteristics are reported in Table 1, with NZ population statistics for comparison purposes. The sub-sample is generally representative of the NZ population. The sub-sample's self-assessed health using the EQ-5D-5L is summarised in Appendix 3.

TABLE 1 Socio-demographic characteristics of the sub-sample (n=2468), and for the NZ population

Characteristic	**Participants**		**NZ population**
	n	**%**	**%[+]**
Age (years)			
18-24	252	10.2	12.8
25-34	439	17.8	16.1
35-44	440	17.8	17.9
45-54	381	15.4	18.8
55-64	402	16.3	15.4
65+	554	22.5	19.0
Gender			
Male	1157	46.9	47.9
Female	1306	52.9	52.1
Gender diverse	5	0.2	unrecorded
Ethnicity*			
New Zealand European	1563	63.3	64.3
Māori	391	15.8	14.1
Pacific	96	3.9	6.9
Asian	239	9.7	11.1
Other	414	16.7	13.6
Education		18+ yrs	15+ yrs
No qualifications/Secondary school	834	33.8	54.2
University degree or equivalent	1079	43.7	20.1
Other post-secondary school qualification	555	22.5	25.7
Individual Income		18+ yrs	15+ yrs
$20,000 or less	552	22.4	38.2
$20,001 - $30,000	425	17.2	13.7
$30,001 - $50,000	537	21.7	21.4
$50,001 - $70,000	414	16.8	12.9
$70,001 - $100,000	336	13.6	7.8
$100,001 or more	204	8.3	6.0
Economic Activity			
Full-time work for pay (30 hours or more per week)	965	39.1[#]	
Part-time work for pay (< 30 hours per week)	380	15.4	
Not in paid work	225	9.1	
Student/Homemaker	329	13.3	
Retired	491	19.9	
Other (including self-employed)	78	3.2	
Long-term Disability (lasting 6 months or more)			
Yes	634	25.7	24.0
No	1834	74.3	76.0

Notes: * Sums to >100% as people identify with multiple ethnicities; [+] Statistics from the NZ 2013 Census; # 2017 September quarter employment and unemployment rates were 67.8% and 4.5%.

Participants answered 20 DCE questions on average, with a median time of 12.6 seconds per answer. As reported in Table 2, most participants (63.7%) found the survey instructions and design easy to understand, and almost half found the DCE questions difficult to answer. For most participants (87%), their ranking of the five dimensions produced by the DCE was as they expected. However, of the remainder (13%), usually either just one dimension was mis-ranked or two dimensions were reversed (often despite equal weights so that if ordinals such as “1st=” had been used, the ranking may have been correct).

TABLE 2 Participant feedback about the DCE

Feedback	Sub-sample (n=2468)	
Understanding instructions/survey design	No.	%
Very easy/Easy	1573	63.7
Neutral	668	27.1
Very difficult/Difficult	227	9.2
Choosing between two health states		
Very Easy/Easy	651	26.4
Neutral	605	24.5
Very Difficult/Difficult	1212	49.1
Ranking of the EQ-5D-5L dimensions		
As expected	2148	87.0
Not as expected	320	13.0

The sub-sample’s mean weights from the DCE are reported in the second column of Table 3, where they are normalised so that level 1 on all dimensions (i.e. ‘full health’) sums to unity: i.e. 11111 = 1. The disutility coefficients are in the third column – where the social value set can be calculated by adding unity to the sum of the coefficients corresponding to each state’s combination of levels; e.g. 55555’s value: 1 – 0.350 – 0.370 – 0.340 – 0.381 – 0.389 = −0.830, 33333’s value: 1 – 0.129 – 0.145 – 0.117 – 0.130 – 0.157 = 0.322, etc. Analogous calculations for all 3125 health states results in a social value set for NZ. This value set – available on the request from the authors – is summarised graphically in Figure 5, where 780 (25%) of the 3125 states are worse than dead = 0.

As a robustness check, additional participant exclusions based on increasing the ‘implausibly quickly’ threshold to median times of 8 and 10 seconds per answer respectively were applied to obtain two smaller, arguably higher-quality, sub-samples of n=2189 and n=1657. Relative to the original n=2468 sub-sample, the social value sets from these two smaller sub-samples are very similar: all three are almost perfectly linearly correlated with each other – their Pearson *r* correlation coefficients all equal 1.000 – and their mean values for dead are 0.338, 0.337 and 0.333. These results suggest that more stringent ‘data quality’ exclusions would be unlikely to significantly alter the social value set.

With reference to the second column in Table 3, the most important dimension for participants on average is Anxiety/Depression with a mean weight of 0.215, followed by Pain/Discomfort (0.206), Self-Care (0.202), Mobility (0.191) and – least important – Usual Activities (0.186). This result for Anxiety/Depression, for example, means that participants value *not* being anxious/depressed the most; in other words, being extremely anxious/depressed confers the most disutility.

With respect to their functional ‘shape’ in terms of marginal effects of level changes, all dimensions exhibit increasing disutility as the levels progressively worsen from “no problems” (level 1) to “moderate problems” (level 3) to “extreme problems” (level 5). Recall that the weights for levels 2 and 4 are interpolated instead of directly created by the DCE – resulting, as can be seen in Table 3, in increasing disutility across all five levels for all dimensions.

TABLE 3 Mean DCE weights and social disutility coefficients (n=2468)

Dimension	**Mean DCE weight**	**Social disutility coefficient**
Mobility		
I have no problems in walking about	0.191	0.000
I have slight problems in waking about	0.159	−0.056
I have moderate problems in walking about	0.119	−0.129
I have moderate problems in walking about	0.065	−0.229
I am unable to walk about	0.000	−0.350
Self-Care		
I have no problems washing or dressing myself	0.202	0.000
I have slight problems washing or dressing myself	0.165	−0.066
I have moderate problems washing or dressing myself	0.121	−0.145
I have severe problems washing or dressing myself	0.065	−0.249
I am unable to wash or dress myself	0.000	−0.370
Usual Activities (e.g. work, study, housework, family or leisure activities)		
I have no problems doing my usual activities	0.186	0.000
I have slight problems doing my usual activities	0.158	−0.050
I have moderate problems doing my usual activities	0.120	−0.117
I have moderate problems doing my usual activities	0.066	−0.217
I am unable to do my usual activities	0.000	−0.340
Pain/Discomfort		
I have no pain or discomfort	0.206	0.000
I have slight pain or discomfort	0.175	−0.055
I have moderate pain or discomfort	0.133	−0.130
I have severe pain or discomfort	0.073	−0.242
I have extreme pain or discomfort	0.000	−0.381
Anxiety/Depression		
I am not anxious or depressed	0.215	0.000
I am slightly anxious or depressed	0.174	−0.072
I am moderately anxious or depressed	0.126	−0.157
I am severely anxious or depressed	0.067	−0.265
I am extremely anxious or depressed	0.000	−0.389

FIGURE 5 The 3125 health state values (means), from highest (11111=1) to lowest (55555=−0.830)

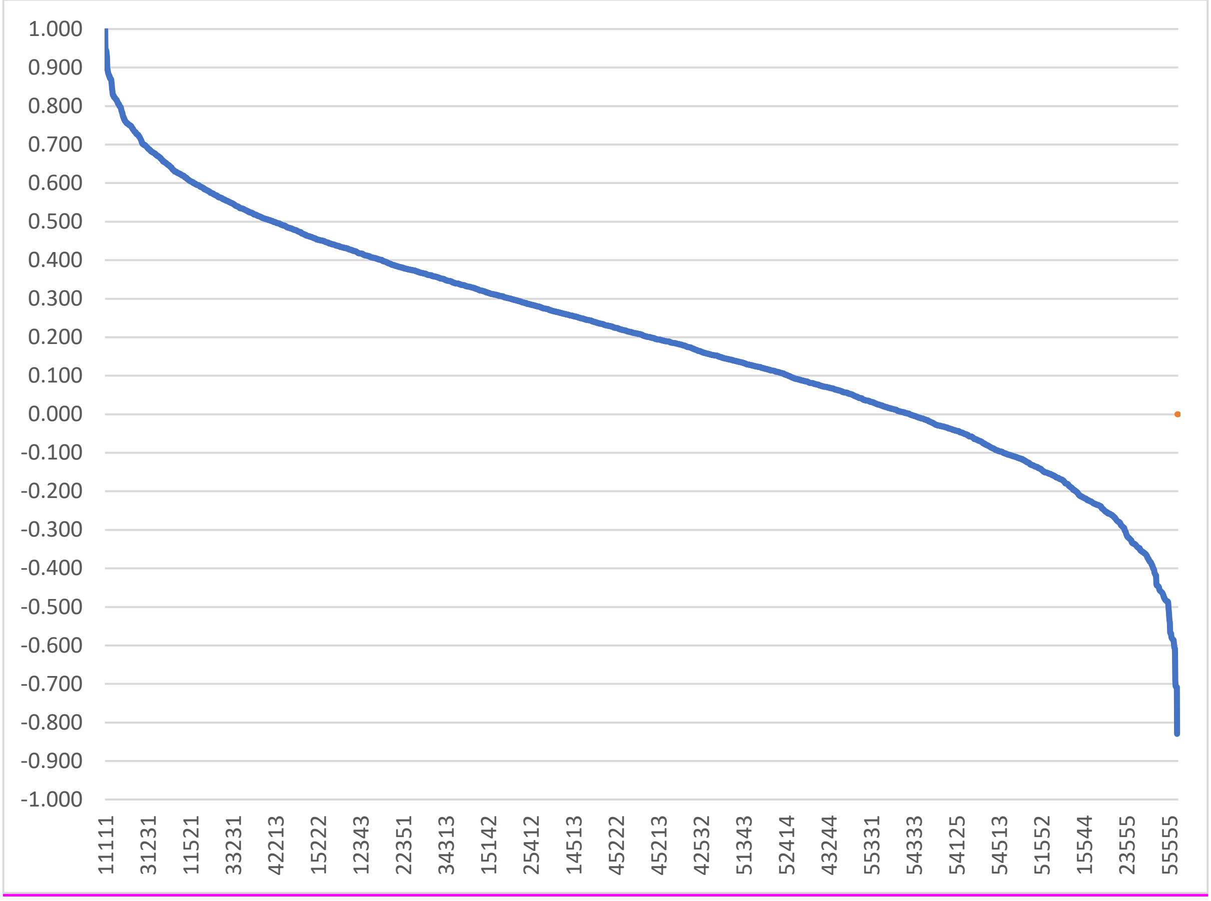


Table 4 reports correlations of the NZ value set vis-à-vis other countries' value sets, their values for 55555 and the number of states worse than dead. All coefficients are statistically significant, except compared with the Netherlands; and NZ has more states worse than dead than the other countries.

TABLE 4 Correlation of the NZ value set with other countries' value sets (EQ-5D-5L TTO)

	Spearman ρ (significance)	Pearson r (significance)	55555 value	No. states worse than dead
Canada	0.957 (0.00)	0.962 (0.00)	−0.284	131
China	0.970 (0.00)	0.973 (0.00)	−0.391	315
Spain	0.946 (0.00)	0.951 (0.00)	−0.416	258
South Korea	0.941 (0.01)	0.945 (0.01)	−0.066	4
Japan	0.970 (0.00)	0.972 (0.00)	−0.025	1
The Netherlands	0.903 (0.09)	0.912 (0.08)	−0.446	483
United Kingdom	0.947 (0.00)	0.951 (0.00)	−0.281	153
Uruguay	0.949 (0.00)	0.954 (0.00)	−0.264	27
New Zealand	- -	- -	−0.830	780

4. DISCUSSION

Fundamental to the new valuation tool is the extensive data quality checks to identify a high-quality sub-sample (n=2468) for creating the social value set (from N=5112 who completed the survey). Participants who answered their DCE questions implausibly quickly or failed to answer the two repeated questions consistently were excluded (as were people whose feedback indicated their answers were invalid, or who valued dead at unity). The social value set created for NZ is robust to more stringent exclusions from increasing the time threshold for identifying people answering implausibly quickly.

On the face of it, excluding more than half – i.e. 2644 (51.6%) – of the people who completed the survey constitutes a high rate of data wastage. However, a low level of participant engagement or understanding is a risk with any survey of the general population, but especially so when recruitment is via an online commercial research panel (as for this study). Fortunately, the tool's data quality checks were able to identify – and exclude – such 'low-quality' participants. In future applications of the tool, recruiting more highly motivated (or incentivised) participants should reduce the need for exclusions.

Nonetheless, almost half of the high-quality sub-sample reported finding the DCE questions difficult to answer. This finding is unsurprisingly given the questions involve confronting trade-offs between EQ-5D-5L dimensions. However, an important advantage of the PAPRIKA method relative to other DCE approaches such as in the EuroQol Group's EQ-VT protocol is that PAPRIKA's questions involve *partial* profiles (just two dimensions at a time) instead of *full* profiles (all five dimensions at a time). Choosing one from two alternative health states, differing on just two dimensions at a time, is the cognitively least complex of *all* possible choice tasks. In future applications of the tool, the inclusion of a preliminary warm-up exercise involving some practice DCE questions, without affecting the person's results, is likely to reduce their initial shock when confronting trade-offs, thereby improving the quality of their data.

An important area for future research relates to how the software determines the weights for levels 2 and 4 for each dimension. In this study, they were interpolated, with only levels 1, 3 and 5 included in the DCE, thereby limiting the number of DCE questions to just 20 on average, typically taking 5-10 minutes in total. Not including level 4 in the DCE also side-steps the confusion that may be experienced by some people when asked to differentiate between levels 4 and 5 on the Pain/Discomfort and Anxiety/Depression dimensions with respect to which of the near-synonyms "severe" versus "extreme" problems is worse. In future applications of the tool, all five levels could easily be included in the DCE – resulting in approximately 55 DCE questions on average (e.g. taking 12-25 minutes in total). Such an elicitation burden is likely to be acceptable for people who are sufficiently motivated (as noted earlier).

Alternatively, an additional DCE could determine mean weights for levels 2 and 4 for each dimension and then be combined with the current results to refine the NZ social value set. Either approach would also enable a proper investigation of each dimension's functional 'shape' in terms of marginal effects of changes across their levels; e.g. if the conjecture above that "severe" and "extreme" problems are near-synonyms then the disutility of going from level 4 to 5 on these dimensions should be minimal.

Feedback overall – from the think-aloud sessions, the pilot-testing and population survey – indicates the tool is acceptable to participants. Its user-friendliness and online delivery could significantly lower the cost of creating EQ-5D-5L value sets. The tool could also support CUA and PROMs being implemented at the individual patient level, incorporating the patient's preferences into their treatment decisions in 'real time'. For example, the tool could be available on computer tablets in doctor waiting rooms or as a mobile app for patients to quickly create their own personal value sets.

From a future research perspective, the availability of personal value sets enables 'clusters' of participants with similar preference weights to be identified using cluster analysis (Späth, 1980) – and whether such clusters vary systematically with their observable characteristics, such as age, ethnicity, employment, health status, etc to be investigated. Also, the large number of health states worse than dead in the NZ social value set created here relative to other countries' value sets (mostly created using the EQ-VT protocol) could be further investigated. Are New Zealanders' preferences fundamentally different from other people's? Or is this result due to the way dead is valued in this study relative to the complexity of the EQ-VT? Or might such personal questions be answered more honestly in private with the new tool than in the presence of an EQ-VT interviewer?

Finally, possible areas for future research internationally include trialling the new tool in other countries – including leveraging the tool's cost advantages for low and middle-income countries – and testing the tool against other methods for creating EQ-5D-5L value sets. An obvious potential comparator is the EQ-VT protocol. The tool could also be adapted to create value sets for other health descriptive systems, such as the SF-6D (Brazier, Roberts, & Deverill, 2002), etc.

5. CONCLUSION

The new valuation tool for creating *personal* and *social* EQ-5D-5L value sets is feasible and acceptable to participants and enables valuation data to be relatively easily and cheaply collected. The tool was first applied in NZ. It could also be used in other countries, tested against other methods for creating EQ-5D-5L value sets, applied in personalised medicine and adapted to create value sets for other health descriptive systems.

REFERENCES

Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, 21(2), 271-292.

Brooks, R. (1996). EuroQol: the current state of play. *Health policy*, 37(1), 53-72.

Devlin, N. J., Hansen, P., Kind, P., & Williams, A. (2003). Logical inconsistencies in survey respondents' health state valuations – A methodological challenge for estimating social tariffs. *Health Economics*, 12(7), 529-544.

Devlin, N., Buckingham, K., Shah, K., Tsuchiya, A., Tilling, C., Wilkinson, G., & Van Hout, B. (2013). A comparison of alternative variants of the lead and lag time TTO. *Health Economics*, 22(5), 517-532.

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 Economics*, 27(1), 7-22.

Devlin, N. J., Shah, K. K., Mulhern, B. J., Pantiri, K., & van Hout, B. (2018). A new method for valuing health: directly eliciting personal utility functions. *The European Journal of Health Economics*, 1-14.

EuroQol. (2018). EQ-5D-5L Valuation. Retrieved 26 July 2018 https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/valuation.

Farin, G., Hoschek, J., & Kim, M.-S. (2002). *Handbook of Computer Aided Geometric Design*. North Holland: Elsevier Science Ltd.

Feng, Y., Devlin, N. J., Shah, K. K., Mulhern, B., & van Hout, B. (2018). New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Economics*, 27(1), 23-38.

Fitzgerald, A., de Coster, C., McMillan, S., Naden, R., Armstrong, F., Barber, A., . . . Lacaille, D. (2011). Relative urgency for referral from primary care to rheumatologists: the Priority Referral Score. *Arthritis Care & Research*, 63(2), 231-239.

French, S. D., Bennell, K. L., Nicolson, P. J., Hodges, P. W., Dobson, F. L., & Hinman, R. S. (2015). What do people with knee or hip osteoarthritis need to know? An international consensus list of essential statements for osteoarthritis. *Arthritis Care & Research*, 67(6), 809-816.

Gilhooly, K., & Green, C. (1996). Protocol analysis: theoretical background. *Handbook of Qualitative Research Methods for Psychology and the Social Sciences*, 14, 43-54. Leicester: The British Psychological Society.

Hansen, P., & Ombler, F. (2008). A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *Journal of Multi-Criteria Decision Analysis*, 15(3-4), 87-107.

Hansen, P., Hendry, A., Naden, R., Ombler, F., & Stewart, R. (2012). A new process for creating points systems for prioritising patients for elective health services. *Clinical Governance: An International Journal*, 17(3), 200-209.

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., . . . Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, 20(10), 1727-1736.

Hernández-Alava, M., Pudney, S., Wailoo, A. (2018). *Quality Review of a Proposed EQ-5D-5L Value Set for England*. Retrieved 10 December 2018 from Universities of Sheffield and York.

Kim, S.-H., Ahn, J., Ock, M., Shin, S., Park, J., Luo, N., & Jo, M.-W. (2016). The EQ-5D-5L valuation study in Korea. *Quality of Life Research*, 25(7), 1845-1852.

Krabbe, P. F., Devlin, N. J., Stolk, E. A., Shah, K. K., Oppe, M., Van Hout, B., . . . Xie, F. (2014). Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. *Medical Care*, 52(11), 935.

Krantz, D. H. (1972). Measurement structures and psychological laws. *Science*, 175(4029), 1427-1435.

Ludwig, K., von der Schulenburg, J.-M. G., & Greiner, W. (2018). German value set for the EQ-5D-5L. *Pharmacoeconomics*, 36(6), 663-674.

Martelli, N., Hansen, P., van den Brink, H., Boudard, A., Cordonnier, A.-L., Devaux, C., . . . Borget, I. (2016). Combining multi-criteria decision analysis and mini-health technology assessment: A funding decision-support tool for medical devices in a university hospital setting. *Journal of Biomedical Informatics*, 59, 201-208.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed), *Frontiers in Econometrics*. Cambridge, MA: Academic Press.

Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for precision medicine. *New England Journal of Medicine*, 366(6), 489-491.

Mulhern, B., Bansback, N., Hole, A. R., & Tsuchiya, A. (2017). Using discrete choice experiments with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. *Medical Decision Making*, 37(3), 285-297.

Norman, R., Cronin, P., & Viney, R. (2013). A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied Health Economics and Health Policy*, 11(3), 287-298.

Ombler, F., Albert, M., & Hansen, P. (2018). How significant are "high" correlations between EQ-5D value sets?. *Medical Decision Making*, 38(6), 635-645.

Oppe, M., Szende, A., & de Charro, F. (2007). Comparative review of Visual Analogue Scale value sets. In: Szende, A., Oppe, M., & Devlin, N.J (Eds), *EQ-5D Value Sets: Inventory, Comparative Review and User Guide*. Berlin, Germany: Springer.

Oppe, M., Devlin, N. J., van Hout, B., Krabbe, P. F., & de Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*, 17(4), 445-453.

Perkins, M. R. V., Devlin, N. J., & Hansen, P. (2004). The validity and reliability of EQ-5D health state valuations in a survey of Māori. *Quality of Life Research*, 13(1), 271-274.

PHARMAC. (2017). *Prescription for Pharmacoeconomic Analysis* (PFPA). Retrieved 10 December 2018: www.pharmac.govt.nz/medicines/how-medicines-are-funded/economic-analysis/pfpa.

Ramos-Goñi, J. M., Pinto-Prades, J. L., Oppe, M., Cabasés, J. M., Serrano-Aguilar, P., & Rivero-Arias, O. (2017). Valuation and modeling of EQ-5D-5L health states using a hybrid approach. *Medical Care*, 55(7), e51-e58.

Richardson, J., McKie, J., & Bariola, E. (2014). Multiattribute utility instruments and their use. In: *Encyclopedia of Health Economics* [Internet]. Elsevier [cited 2017 Feb 5]. p. 341–357. Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780123756787005058.

Robinson, A., & Spencer, A. (2006). Exploring challenges to TTO utilities: Valuing states worse than dead. *Health Economics*, 15(4), 393-402.

Shiboski, C. H., Shiboski, S. C., Seror, R., Criswell, L. A., Labetoulle, M., Lietman, T. M., . . . Bowman, S. J. (2017). 2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome: A consensus and data-driven methodology involving three international patient cohorts. *Arthritis & Rheumatology*, 69(1), 35-45.

Shiroiwa, T., Ikeda, S., Noto, S., Igarashi, A., Fukuda, T., Saito, S., & Shimozuma, K. (2016). Comparison of value set based on DCE and/or TTO data: Scoring for EQ-5D-5L health states in Japan. *Value in Health*, 19(5), 648-654.

Shuker, C., Bohm, G., Hamblin, R., Simpson, A., St George, D., Stolarek, I., Wilson, J., & Merry, A. F. (2017). Progress in public reporting in New Zealand since the Ombudsman's ruling, and an invitation. *The New Zealand Medical Journal* (Online), 130(1457), 11-22.

Späth, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Chichester, UK: Ellis Horwood.

Stolk, E., Ludwig, K., Rand, K., van Hout, B., & Ramos-Goñi, J. M. (2019). Overview, update, and lessons learned from the international EQ-5D-5L valuation work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health*. In Press. Available from: https://doi.org/10.1016/j.jval.2018.05.010.

Sullivan, T., & Hansen, P. (2017). Determining criteria and weights for prioritizing health technologies based on the preferences of the general population: A New Zealand pilot study. *Value in Health*, 20(4), 679-686.

Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., . . . Carmeli, Y. (2018). Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*, 18(3), 318-327.

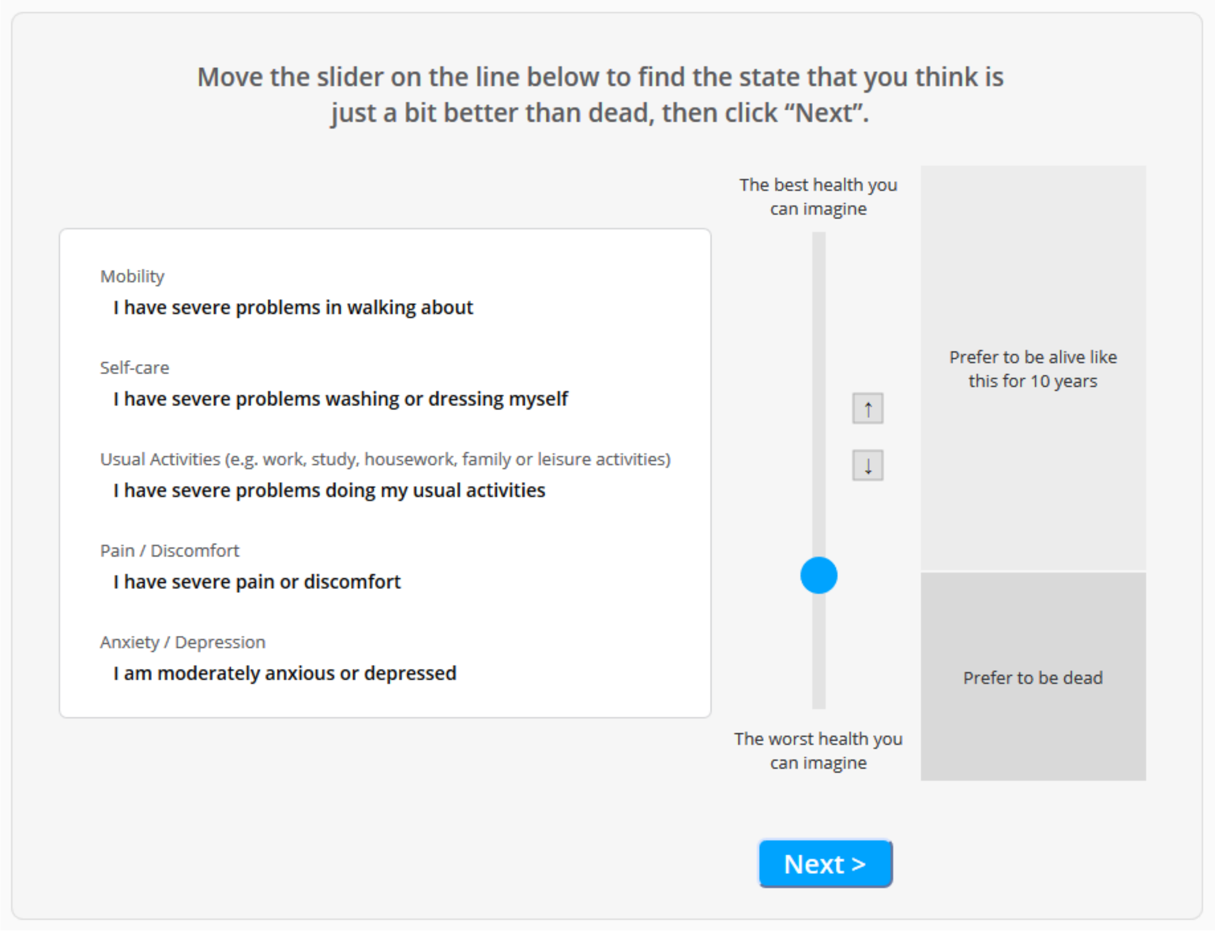
Versteegh, M. M., Vermeulen, K. M., Evers, S. M., de Wit, G. A., Prenger, R., & Stolk, E. A. (2016). Dutch tariff for the five-level version of EQ-5D. *Value in Health*, 19(4), 343-352.

Viney, R., Norman, R., Brazier, J., Cronin, P., King, M. T., Ratcliffe, J., & Street, D. (2014). An Australian discrete choice experiment to value EQ-5D health states. *Health Economics*, 23(6), 729-742.

APPENDIX 1

Below are two alternative user interfaces for identifying health states worse than dead; both were superseded by the configuration presented in Figure 2.

FIGURE A1 Example of a VAS question to identify states worse than dead



Move the slider on the line below to find the state that you think is just a bit better than dead, then click "Next".

Mobility
I have severe problems in walking about

Self-care
I have severe problems washing or dressing myself

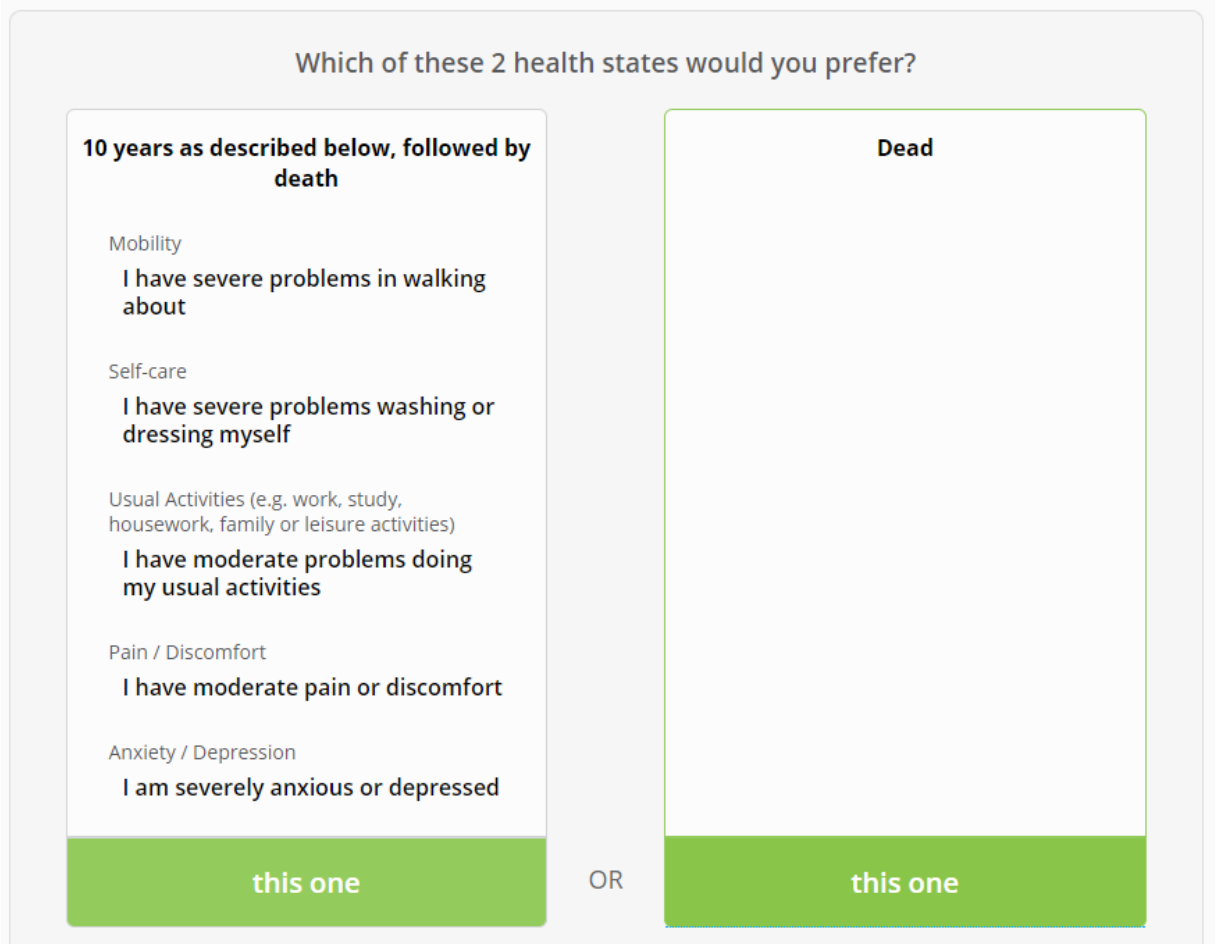
Usual Activities (e.g. work, study, housework, family or leisure activities)
I have severe problems doing my usual activities

Pain / Discomfort
I have severe pain or discomfort

Anxiety / Depression
I am moderately anxious or depressed

The best health you can imagine
↑
↓
The worst health you can imagine
Prefer to be alive like this for 10 years
Prefer to be dead
Next >

FIGURE A2 Example of a pairwise-ranking question to identify states worse than dead



Which of these 2 health states would you prefer?

10 years as described below, followed by death

Mobility
I have severe problems in walking about

Self-care
I have severe problems washing or dressing myself

Usual Activities (e.g. work, study, housework, family or leisure activities)
I have moderate problems doing my usual activities

Pain / Discomfort
I have moderate pain or discomfort

Anxiety / Depression
I am severely anxious or depressed

Dead

APPENDIX 2

Participants' self-assessed health, as described by the EQ-5D-5L, is summarised in Table A1. No NZ EQ-5D-5L population norms are available for comparing these results. Although most participants reported being in good health, almost 50% reported some level of Anxiety/Depression and approximately 62% reported some level of Pain/Discomfort.

TABLE A1 Percentage of sub-sample (n=2468) reporting levels 1-5 on EQ-5D-5L dimensions

Dimension	**Level 1 %**	**Level 2 %**	**Level 3 %**	**Level 4 %**	**Level 5 %**
Mobility	72.1	17.7	7.4	2.3	0.5
Self-care	91.4	6.5	1.5	0.4	0.2
Usual activity	70.2	19.9	6.8	2.7	0.4
Pain/Discomfort	38.3	42.8	14.8	3.5	0.6
Anxiety/Depression	53.6	28.9	12.1	4.1	1.3
EQ VAS	**Mean**	**SD**	**Median**	**25th**	**75th**
	74.8	17.9	74.3	67.5	90