

Complex Valuation: Applying Ideas from the Complex Intervention Framework to Valuation of a New Measure for End-of-Life Care

Joanna Coast¹ · Elisabeth Huynh² · Philip Kinghorn¹ · Terry Flynn³

Published online: 29 December 2015
© Springer International Publishing Switzerland 2015

Abstract

Background The UK Medical Research Council approach to evaluating complex interventions moves through development, feasibility, piloting, evaluation and implementation in an iterative manner. This approach might be useful as a conceptual process underlying complex valuation tasks.

Objective The objective of the study was to explore the applicability of such a framework using a single case study (valuing the ICECAP-Supportive Care Measure) and considering three key uncertainties: the number of response categories for the measure; experimental design; and the potential for using slightly different variants of the measure with the same value set.

Methods Three on-line pilot studies ($n = 204$, $n = 100$, $n = 102$) were undertaken during 2012 and 2013 with adults from the UK general population. Each used variants of discrete choice and best-worst scaling tasks; respondents were randomly allocated to different groups to allow exploration of the number of levels for the instrument (four or five), optimal experimental design and the values for alternative wording around prognosis. Conditional logit regression models were used in the analysis and variance scale factors were explored.

Results The five-level version of the measure seemed to result in simplifying heuristics. Plotting the variance scale

factors suggested that best-worst scaling answers were approximately four times more consistent than the discrete choice answers. The likelihood ratio test indicated there was virtually no difference in values between the differently worded versions.

Conclusion Rigorous piloting can improve the design of valuation studies. Thinking in terms of a ‘complex valuation framework’ may emphasise the importance of conducting and funding such rigorous pilots.

Key Points

The Medical Research Council complex intervention framework is a useful basis for developing and evaluating complex interventions.

The approach could be adapted as a useful way of thinking about developing and piloting valuation tasks, focusing on complex valuation.

The ideas of complex valuation are applied to a valuation task in end-of-life care and show how piloting can influence the design of the valuation task, but also the instrument to be valued and the interpretation of the final values.

✉ Joanna Coast
jo.coast@bristol.ac.uk

¹ School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

² Institute for Choice, University of South Australia, Adelaide, Australia

³ TF Choices Ltd, Nottingham, UK

1 Introduction

Many of those involved in valuation studies are familiar with the two UK Medical Research Council (MRC) frameworks for developing and evaluating complex

interventions. These outline an approach to the process of assessing the effectiveness of interventions, and contain components including “behaviours, parameters of behaviours (e.g. frequency, timing,) and methods of organising and delivering those behaviours (e.g. type(s) of practitioner, setting and location)” [1]. The first framework [1, 2] advocates a linear stepwise process moving through five phases of: theory, modelling, exploratory trial, definitive randomised controlled trial and long-term implementation. The second framework proposes a more circular iterative process through steps of development, feasibility and piloting, evaluation and implementation, but with the possibility of shifts forwards and backwards between each stage [3]. It suggests “a carefully phased approach, starting with a series of pilot studies targeted at each of the key uncertainties in the design, and moving on to an exploratory and then a definitive evaluation” [3].

The broad approach within this second framework could be helpful in valuation studies, which are often complex and where there may be questions about both the development of the instrument and the valuation techniques to be used, with the latter encompassing issues of instrument design, experimental design, the appropriate sample group and the acceptability of the valuation exercise. The importance of adequately conducting and reporting development work prior to instrument valuation has recently been highlighted [4], but similar arguments are applicable to conducting and reporting pilot studies for valuation exercises; piloting is often mentioned in just one or two sentences and is seldom described in detail; thus, it is difficult to discern its impact on the final study design. In contrast, the MRC framework emphasises the importance of using pilot work to “examine key uncertainties that have been identified during development” [3] and highlights the importance of wide dissemination [3]. Not reporting such work for valuation exercises is unhelpful for those conducting later work and casts doubt on the rigour of these exercises.

As with an earlier paper on attribute development [5], this paper provides a case study of one valuation exercise, attempting to be open about the nature of piloting, and how the pilots influenced the final study design. Inevitably, any case study will illuminate specific issues, but this case study was chosen to be helpful in illustrating the notion of ‘complex valuation’: as with the complex intervention framework, the piloting phase had links both backwards to the ‘development’ stage (final design of the instrument) and forwards to the ‘evaluation’ stage (anticipated use of the measure). The case study thus goes beyond what is traditionally considered in piloting, which is often limited to the design of the experiment itself.

The paper begins by presenting the case study setting and then presenting three key areas of uncertainty related to

the design of the measure that were considered during piloting. The value of thinking of these issues in a manner akin to the complex intervention framework is then discussed.

2 The Case Study: ICECAP Supportive Care Measure

The case study reported here concerns the development of an instrument intended for use in economic evaluations of interventions for those at the end of life. For end-of-life care, focusing on health alone may be considered inadequate [6, 7] as the aims of care relate to important outcomes beyond health including maintaining dignity, providing support and enabling preparation for death [8]. This case study focuses on the ICECAP Supportive Care Measure (ICECAP-SCM), designed to be amenable to valuation and subsequent use in economic evaluation [9].

The descriptive system for ICECAP-SCM was generated using informant-led interviews [9, 10] and has seven conceptual attributes: Choice (‘being able to make decisions about my life and care’); Love and affection (‘being able to be with people who care about me’); Physical suffering (‘experiencing significant physical discomfort’); Emotional suffering (‘experiencing emotional suffering’); Dignity (‘being able to maintain my dignity and self-respect’); Being supported (‘being able to have the help and support that I need’); and Preparation (‘having had the opportunity to make the preparations I want to make’) [9]. Examples are given for each attribute to help to clarify meaning. The published descriptive system has four levels ranging from having an attribute ‘none of the time’ through ‘a little of the time’, ‘some of the time’ to ‘most of the time’ [9].

There is a need to develop values for the ICECAP-SCM to enable outcomes from end-of-life care interventions, measured using the descriptive system, to be valued and thus contribute to estimating the relative cost effectiveness of alternative interventions. The overall approach to valuation drew on that of earlier ICECAP measures in using Case 2 (Profile case) best-worst scaling (BWS) [11], but there was also a desire to explore the use of discrete choice experiments (DCEs) [12] to allow the acquisition of additional preference information around interactions [13].

The aim was to pursue the extra-welfarist approach [14, 15] to economic evaluation (as supported by the National Institute for Health and Care Excellence [16]) of obtaining general population values. Whilst there are also arguments for considering values from those at the end of life (because of the unique nature of the experience and the possibility that values shift towards the end of life [17]), it may be difficult to obtain values from these individuals because

of their physical and emotional vulnerability. It is particularly important to know whether values differ between these two groups, and therefore a major aspect of the design task was to generate a design that could, in time, facilitate such a comparison.

2.1 Identification of Key Uncertainties

In line with MRC guidance for evaluating complex interventions, “a series of pilot studies targeted at each of the key uncertainties in the design” [3] was identified and conducted. Key uncertainties related to the design of the measure that was to be valued, the design of the experiment itself and the design of future studies that would use the measure.

2.1.1 Key Uncertainty 1: Design of the Measure

Although the important attributes had been established through detailed qualitative work [4, 9], there was remaining uncertainty about the final design in terms of the number of response categories (‘levels’). The qualitative work suggested that four response categories would be most appropriate because informants did not generally feel that having an attribute ‘all of the time’ was realistic in terms of what could be achieved by and for those at the end of life. Measures used in extra-welfarist approaches to economic evaluation, however, generally have valuations with maximum levels of ‘full health’ or ‘full capability’ and thus there was a question about whether an additional ‘top’ level for each attribute representing ‘all of the time’ should be included.

The number of response categories may, however, influence how respondents complete valuation exercises. It was anticipated that moving from four to five levels could increase the complexity of the valuation task, but might also make it easier for respondents to use simplifying heuristics (choosing the ‘middle’ level). It was therefore important to use this information alongside the theoretical and qualitative information in making a final decision about the number of response categories.

2.1.2 Key Uncertainty 2: Design of the Experiment

Two key uncertainties related to experimental design. The first concerned designing an experiment to maximise the information available from the general population; the second concerned, at the same time, designing a smaller subset to the experiment that would still provide meaningful information and would be feasible for collecting a comparable set of data from those at the end of life.

It was not possible to collect data from those at the end of life during piloting, for ethical and feasibility reasons. Instead, the aim was to identify, through work with the general population, those aspects of the task that would provide comparable data but also be potentially feasible for future use with a vulnerable population group who would find a long and repetitive task too burdensome. As a second stage of the ‘complex valuation process’ (beyond the scope of this paper), the work would go on to directly determine the feasibility of the chosen task amongst people at the end of life.

2.1.3 Key Uncertainty 3: Design of Future Studies Using the Measure

A key difficulty with the conduct of studies at the end of life is that not all respondents are aware of their prognosis, or willing to be faced with a measure that is open about this prognosis. Within ICECAP-SCM, only one attribute, Preparation, is explicit about the end-of-life status of the individual. In the standard version, it is worded as ‘Being prepared—Having financial affairs in order, having your funeral planned, saying goodbye to family and friends, resolving things that are important to you, having treatment preferences in writing or making a living will’ [9]. For some purposes, a less overt reference to end of life is required, and therefore a less explicit wording for the Preparation attribute was developed that omitted reference to funeral arrangements and saying goodbye. A key design issue relating to the use of the measure in future studies was whether values could be used interchangeably between the two versions.

2.2 Methods

2.2.1 Data Collection

A DCE [12] incorporating a Case 2 BWS exercise [11] was used. Three on-line pilot surveys were administered in an iterative manner to an adult general UK population sample. Uncertainty around the design of the measure in terms of the number of response categories was explored in the first and second pilot, uncertainty around the experimental design and cognitive ease was explored throughout the three pilots and uncertainty around future use of the measure in terms of the standard and less explicit questionnaire versions was explored in the third pilot.

Within each pilot, each respondent completed 16 tasks. Each began with a BWS task providing a scenario from which a respondent was asked to indicate the best and worst thing about that end-of-life state, and was followed

Set 1 of 16

Imagine living in the end of life state presented and decide which aspect you think would be **most** acceptable, and which aspect you think **least** acceptable? Please select one answer per column.

	 Most	 Least
I am able to have the help and support that I need some of the time	<input type="radio"/>	<input type="radio"/>
I always experience significant physical discomfort	<input type="radio"/>	<input type="radio"/>
I have not had the opportunity to make any of the preparations I want to make	<input type="radio"/>	<input type="radio"/>
If I want to, I am able to be with people who care about me some of the time	<input type="radio"/>	<input type="radio"/>
I am able make decisions that I need to make about my life and care some of the time	<input type="radio"/>	<input type="radio"/>
I sometimes experience emotional suffering	<input type="radio"/>	<input type="radio"/>
I am never able to maintain my dignity and self-respect	<input type="radio"/>	<input type="radio"/>

Imagine you are able to choose between the current end of life state (A) and a "middling" end of life state (B), which would you choose?

End of life state A	End of life state B
I am able to have the help and support that I need some of the time	I am able to have the help and support that I need some of the time
I always experience significant physical discomfort	I often experience significant physical discomfort
I have not had the opportunity to make any of the preparations I want to make	I have had the opportunity to make some of the preparations I want to make
If I want to, I am able to be with people who care about me some of the time	If I want to, I am able to be with people who care about me only a little of the time
I am able make decisions that I need to make about my life and care some of the time	I am able to make decisions that I need to make about my life and care some of the time
I sometimes experience emotional suffering	I often experience emotional suffering
I am never able to maintain my dignity and self-respect	I am able to maintain my dignity and self-respect some of the time
<input type="radio"/> End of life state A	<input type="radio"/> End of life state B

>>

Fig. 1 Screenshot of choice task taken from pilot 3

by a DCE task with a choice between the state described by the BWS task and a comparator end-of-life state. Figure 1 presents a screenshot of the task respondents completed in pilot 3 (pilots 1 and 2 differ in the levels that were presented). Key sociodemographic data, EQ-5D-5L (measuring health functioning) [18] and ICECAP-A (measuring

capability well-being) [19, 20] were collected as were data on completion time, recorded automatically by the on-line system.

The research was reviewed and approved by the Science, Technology, Engineering and Mathematics Ethical Review Committee of the University of Birmingham

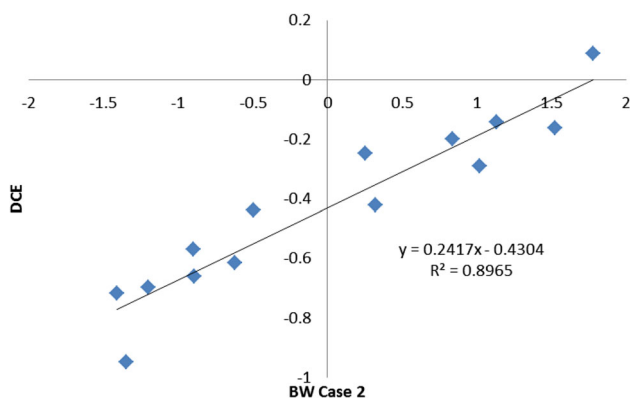


Fig. 2 Plot of best-worst (BW) scaling against a discrete choice experiment (DCE) [pilot 1] showing relative variance scale factors

(ERN_11-1296). After being informed about the work, potential participants were asked to click if they agreed to take part in the survey.

2.2.1.1 Uncertainty Around Measure Design and Number of Levels In pilot 1, respondents were randomly assigned on a 1:1 basis to the four- or five-level version of the instrument. Pilot 2 was conducted to ensure that the findings from pilot 1 did not result from design artefacts associated with the representation of the choice task for the four-level version.

2.2.1.2 Uncertainty Around Experimental Design and Cognitive Ease This aspect of piloting was designed to compare the task efficiency and choice consistency of a variety of linked BWS and DCE tasks. In designing a simplified subset of the task for use with those at the end of life, it was anticipated, a priori, that only a BWS or a DCE task would be administered. One means of assisting the choice between these two options was to consider the consistency of responses to the different tasks: more consistent responses suggest that less effort is required to complete the task, which may be appropriate grounds for identifying a task that is more feasible for those at the end of life. Uncertainty around the overall experimental design was explored iteratively through the pilots. In the first and second pilot, the BWS task presented only the top and bottom levels of each attribute, and was followed by a DCE task with a choice between the state described by the BWS task and a “middling” constant. This was intended to provide preliminary indication of the relative statistical efficiency and choice consistency of the two tasks (BWS and pairwise DCE). An orthogonal main effects plan (OMEP) for the two “extreme” levels (a 2^7 design) required only eight choice sets. Two versions of the OMEP in eight were used, rotating the coding for the second, to test for ordering effects in attributes. Pilot 3 randomised

respondents on a 1:1 basis to one half of the full 4^7 OMEP in 32 choice sets and again used a constant “middling” state for the DCE comparator. Pilot 3 was designed to generate preliminary estimates of intermediate levels for the attributes and further information about choice consistency.

2.2.1.3 Uncertainty Around Future Use of the Measure and Questionnaire Version In pilot 3, respondents were, further, randomly assigned on a 1:1 basis to the standard or less explicit questionnaire version.

2.2.2 Data Analysis

The behavioural model was assumed to be a logit model. Conditional logit regression models [21] were used to estimate part-worth utilities for ICECAP-SCM for the BWS and DCE methods.

2.2.2.1 Uncertainty Around Measure Design and Number of Levels Model estimates from the four- and five-level versions were compared. The analysis looked for systematic or significant effects on the ordering of attribute relative importance between the two versions. Additionally, frequency counts of choices were used to detect potential simplifying decision strategies, with particular focus on whether the version wording appeared to induce respondents to systematically choose a “middling” option in the DCE; a version inducing such simplifying strategies would be of considerable concern.

2.2.2.2 Uncertainty Around Experimental Design and Cognitive Ease The variance scale factor between the BWS and DCE data was obtained by plotting one set of estimates against the other [22]; the variance scale factor is concerned with how consistent individuals are in making their choices and informs the question of cognitive ease. For the experimental design more generally, pilots 1 and 2 provided estimates of the two “extreme” levels; pilot 3 enabled preliminary estimates of intermediate levels. Attribute levels were first dummy coded, then linearity was imposed on the slope coefficients for attributes that were close to linear preferences across levels; linearity was imposed as decrements from the top level for the specific attribute. Following advice (personal communication, John Rose), alternative models were run including interaction terms representing the 21 bottom-by-bottom level interactions (being the most likely to be non-zero given the need for a relatively parsimonious model). OMEPs cannot properly separate interaction effects but some may be estimated (because, for a non-linear model such as the LOGIT model with Gumbel distribution, orthogonality does not hold for the variance covariance matrix making

estimating interactions possible). Interactions were not expected to be significant but were anticipated to offer insights into multiplicative relationships between attributes that would be useful for designing the main survey.

2.2.2.3 Uncertainty Around Future Use of the Measure and Questionnaire Version Separate models were estimated for each questionnaire version. Estimated utilities were plotted to allow observation of systematic effects on the ordering of relative importance for Preparation. A conditional logit model on the pooled data of both versions was then estimated with additional version specific indicators for Preparation. A likelihood ratio test on the version-specific coefficients tested for significant differences in questionnaire versions for Preparation.

3 Results

Three on-line pilot surveys were administered between September 2012 and March 2013 (pilot 1, $n = 204$; pilot 2, $n = 100$; pilot 3, $n = 102$). Table 1 summarises characteristics of survey respondents and Table 2 shows response frequencies for ICECAP-A dimensions.

3.1 Uncertainty Around Measure Design and Number of Levels

Separate analysis of the two OMEPs (each in eight choice sets, with different coding used) established there were no order effects. Attributes with particularly attractive (unattractive) upper (lower) levels had higher estimated “best” (“worst”) values in both OMEPs. As expected, owing to small numbers of choices (low precision), differences between the OMEPs were only observed in the best (worst) choice data for top (bottom) levels. Love and affection, Choice and Dignity were the top three attributes in both OMEPs in the BWS exercise and in the four-level version of the DCE, but Physical suffering displaced Choice in the five-level DCE version.

The five-level version data were, however, somewhat unreliable. The most consistent difference between the four- and five-level versions was evident in the DCE task. For the four-level version ($n = 102$), six individuals always chose the middling state (state 3232323, states ranging from 1111111 to 4444444), whereas with the five-level version ($n = 102$), 29 individuals always chose the middling state (state 3333333, states ranging from 1111111 to 5555555). The five-level version seemed to induce respondents to stick to the “middling” option; a version that induces such simplifying strategies is problematic. Pilot 2 changed the choice of “middling” state, using 3232323 instead of 2323232, and found similar results, suggesting that the findings were not

Table 1 Sample characteristics for each pilot (proportions, to one decimal place, except for age)

	Pilot 1 ($n = 204$)	Pilot 2 ($n = 100$)	Pilot 3 ($n = 102$)
Sex (%)			
Male	45.1	48	41.2
Ethnicity (%)			
White (British or other)	89.7	89	92.2
Mixed/multiple ethnic groups	1.0	1	2.9
Asian/Asian British	7.4	5	3.9
Black/African/Caribbean/Black British	2.0	2	1.0
Any other ethnic group	0	3	0
Relationship (%)			
Married/civil partnership	47.1	53	57.8
In a relationship (not married)	23.5	21	17.6
Single	27.9	24	24.5
Widowed/surviving partner of a civil partnership	1.5	2	0
Bereaved within the previous 2 years (%)	38.2	37	29.4
Religion (%)			
No religion	39.7	40	41.2
Christian	49.5	55	49.0
Buddhist	0.5	1	1.0
Hindu	1.0	0	0
Jewish	2.0	1	1.0
Muslim	5.9	2	5.9
Other	1.5	1	2.0
Employment (%)			
Employed (full or part time)	68.6	64	63.7
Self-employed	8.3	7	4.9
Looking after the home and/or family	4.4	7	5.9
Unemployed	4.9	8	4.9
Retired	4.4	8	14.7
Full-time education	5.9	3	2.0
Long-term sick leave/disabled	2.9	3	2.9
Full-time carer	0.5	0	1.0
Age (years)			
Mean	39.4	39.9	39.9
Standard deviation	12.0	11.9	11.9
Minimum	18	18	18
Maximum	77	71	71

peculiar to the choice of middling state in the first pilot. It appeared that the “variation” in the middling states (levels 2 or 3) in the four-level version caused respondents to pay greater attention to the levels offered.

Table 2 ICECAP-A response frequencies (%), by pilot

	Pilot 1 (<i>n</i> = 204)	Pilot 2 (<i>n</i> = 100)	Pilot 3 (<i>n</i> = 102)
Feeling settled and secure			
I am able to feel settled and secure in all areas of my life	32.8	33	35.3
I am able to feel settled and secure in many areas of my life	47.1	38	49
I am able to feel settled and secure in a few areas of my life	15.2	19	14.7
I am unable to feel settled and secure in any areas of my life	4.9	10	1
Love, friendship and support			
I can have a lot of love, friendship and support	47.1	54	49
I can have quite a lot of love, friendship and support	36.3	30	34.3
I can have a little love, friendship and support	14.7	15	12.7
I cannot have any love, friendship and support	2	1	3.9
Being independent			
I am able to be completely independent	65.2	56	69.6
I am able to be independent in many things	26.5	31	23.5
I am able to be independent in a few things	6.9	12	6.9
I am unable to be at all independent	1.5	1	0
Achievement and progress			
I can achieve and progress in all aspects of my life	35.3	42	40.2
I can achieve and progress in many aspects of my life	48	33	44.1
I can achieve and progress in a few aspects of my life	13.2	19	13.7
I cannot achieve and progress in any aspects of my life	3.4	6	2
Enjoyment and pleasure			
I can have a lot of enjoyment and pleasure	45.1	48	44.1
I can have quite a lot of enjoyment and pleasure	37.7	35	43.1
I can have a little enjoyment and pleasure	15.7	13	9.8
I cannot have any enjoyment and pleasure	1.5	4	2.9

3.2 Uncertainty Around Experimental Design and Cognitive Ease

Plotting the BWS and DCE estimates showed a large relative variance scale factor (smaller error variance) for the BWS and suggested that BWS answers were approximately four times more consistent than those from the DCE (Fig. 2). This implies that, for similar effort, more predictive power is achieved through the BWS task and that this should be the focus for the simplified patient task.

In terms of information for generating the overall experimental design, pilot 3 results indicated that Dignity, Physical suffering and Love and affection made the largest contributions to total value. BWS results suggested decrements were broadly linear by level. Table 3 presents the slope coefficients from the DCE results after imposing linearity. Preparation was not significant in any pilot. Table 4 presents coefficients for the estimable interaction terms. Post-hoc justification is hazardous but members of the project team indicated that the signs of these terms were sensible. Although Preparation was not important as a main effect, it may be important when accompanied by other impairments.

3.3 Uncertainty Around Future Use of the Measure and Questionnaire version

The likelihood ratio test on the version specific coefficients ($LL(U) = -928.122$; $LL(R) = -929.884$; $LRT = 3.5238$ (3 degrees of freedom); $P(LRT) = 0.682$) indicated there were virtually no differences in part-worth utilities associated with the two questionnaire versions.

3.4 Implications of the ‘Complex Valuation’ Pilot Work for Final Study Design

3.4.1 Measure Design and Number of Levels

The pilots established that the five-level version of ICECAP-SCM seemed to result in the use of simplifying heuristics, and that the four-level version did not suffer from such issues. Given earlier evidence from the qualitative work that also supported the four-level version, this version of ICECAP-SCM was taken forward.

Table 3 Discrete choice experiment conditional logit model results imposing linearity (pilot 3, $n = 100$)

	Coefficient	Standard error ^a	95 % confidence interval	
Being supported	-0.2289*	0.0504	-0.3276	-0.1301
Physical suffering	-0.4274*	0.0629	-0.5506	-0.3042
Preparation	-0.0540	0.0467	-0.1456	0.0375
Love and affection	-0.3645*	0.0604	-0.4828	-0.2462
Choice	-0.17701*	0.0507	-0.2764	-0.0777
Emotional suffering	-0.2794*	0.0490	-0.3754	-0.1833
Dignity	-0.4580*	0.0611	-0.5778	-0.3383
Log pseudolikelihood	-943.4162			
Pseudo R2	0.1407			
Wald chi2(8)	93.21			

* Significant at the 10, 5 and 1 % level

^a Robust standard errors to account for clustering at the respondent level**Table 4** Discrete choice experiment conditional logit model interaction effects (pilot 3, $n = 100$)

Interaction variables	Coefficient	Standard error ^a	95 % confidence interval	
Being supported × physical suffering	1.0305	1.1276	-1.1796	3.2406
Being supported × preparation	-0.2917	0.8519	-1.9613	1.3779
Being supported × love and affection	0.8526*	0.4845	-0.0971	1.8023
Being supported × choice	-0.6658	0.4483	-1.5444	0.2128
Being supported × emotional suffering	-0.0047	0.4824	-0.9502	0.9408
Being supported × dignity	-0.1342	0.4183	-0.9540	0.6856
Physical suffering × preparation	-1.175*	0.6917	-2.5307	0.1808
Physical suffering × love and affection	1.1084*	0.5757	-0.0200	2.2368
Physical suffering × choice	1.5885**	0.6266	0.3603	2.8166
Physical suffering × emotional suffering	0.0349	0.5386	-1.0207	1.0905
Physical suffering × dignity	-0.2982	0.5078	-1.2936	0.6971
Preparation × love and affection	-2.249**	1.0308	-4.2691	-0.229
Preparation × choice	0.0072	0.3938	-0.7647	0.7791
Preparation × emotional suffering	0.5827	0.6930	-0.7756	1.9410
Preparation × dignity	1.6569**	0.7170	0.2516	3.0623
Love and affection × choice		(Omitted)		
Love and affection × emotional suffering	0.6681	0.6972	-0.6984	2.0347
Love and affection × dignity	0.7747	0.6289	-0.4579	2.0073
Choice × emotional suffering	-0.739	0.5491	-1.8152	0.3372
Choice × dignity		(Omitted)		
Emotional suffering × dignity	-0.0249	0.4786	-0.9629	0.9131
Log pseudolikelihood	-925.375			
Pseudo R2	0.1572			
Wald chi2(27)	128.22***			

Estimates for the main effects are not shown, available on request

* ** *** Significance at the 10, 5 and 1 % level

^a Robust standard errors to account for clustering at the respondent level

3.4.2 Experimental Design and Cognitive Ease

Estimates from Table 3, together with estimates of main effects (available on request), were used to generate an efficient design for the main general population survey.

The identification of heterogeneity in the piloting was inevitably limited so, to the extent that mean values from pilot 3 are incorrect for some population segments, the design will not be optimal, resulting in larger standard errors. A large planned sample size for the full survey

($n = 6000$), however, means such effects will be reduced. The final design for the main valuation exercise comprised eight common sets (to enable robust heterogeneity analysis) and five blocks of eight (to estimate two-way interactions). This design should enable the generation of a tariff of general population values, based on main effects and two-way interactions. It will also allow characterisation of “types” of respondent.

Answers from the BWS appeared to be much more consistent than those from the DCE (reflecting previous work [23]). This suggests BWS as the appropriate format for future work to assess the feasibility of eliciting values from those at the end of life. Further, completion of just the common eight BWS choice sets provides sufficient information to allow estimation of main effects, meaning that a much simplified task, requiring the consideration of just eight scenarios rather than the 32 considered in the general population task, is possible. If this task proves feasible, it will, in time, be possible to obtain a core comparable set of data from those at the end of life, enabling general population values to be set in the relevant context of the values directly held by those receiving end-of-life care.

3.4.3 Future Use of the Measure

Finally, the standard and less explicit versions of ICECAP-SCM appeared not to be associated with differences in values, and so it was concluded that values obtained using the standard version of the questionnaire could be used interchangeably across both versions. The terminology used in the standard version of the questionnaire was taken forward for use in the main valuation exercise.

4 Discussion

As with designing complex interventions, there are many ways in which pilot findings can inform complex valuation tasks. These include the usual consideration of experimental and statistical design, and more novel and nuanced questions relating to earlier and later stages in the design such as those explored here in terms of the number of levels for the measure and whether slightly different variants of the measure can be used. Researchers considering piloting for valuation studies could find more explicit consideration of the types of factors included in the complex intervention framework helpful in expanding the possibilities for learning from pilot data.

This paper described one case study exploring the development of rigorous piloting for generating a design to obtain robust values for the ICECAP-SCM. The case study concentrated on eliminating or reducing key uncertainties

associated with designing both the instrument and the valuation task. As with the newer version of the MRC complex intervention framework [3], the feasibility piloting had both ‘backward’ influences on the instrument descriptive system and ‘forward’ influences on the design for the valuation study. The iterative nature of designing complex interventions (here, complex valuation) was also evident, with the introduction of pilot 2 following results obtained in the first pilot. Ongoing future iterations will also determine whether, in practice, interview administering a BWS task with those at the end of life is feasible.

This research also points to areas that might be helpful for those piloting future studies in a ‘complex valuation’ framework. In relation to discrete choice, there is growing realisation that neither conventional DCEs nor Case 2 BWS offers unequivocal benefits [24]. Other studies may also be able to draw upon the ability of DCEs to allow estimation of interactions and the ability of BWS to estimate attribute levels more precisely than a DCE. More generally, the novel “nesting” structure of the task designed here should be considered for future valuation exercises that require inclusion of vulnerable individuals.

Further research is required to develop a ‘complex valuation’ framework. This paper has outlined the possibility of developing a framework similar to that for the evaluation of complex interventions. It may, of course, be that the most rigorous research on valuation is already working along similar lines. The advantage of a more explicit framework would be to normalise this type of approach amongst researchers, but also amongst research funders. It is notable, for example, that the MRC framework suggests that “researchers should be prepared to explain to decision makers the need for adequate development work” [3] and a similar ‘complex valuation’ framework, which advocates for the resources for adequate piloting, may be helpful in generating more robust research findings.

5 Conclusion

This work shows the benefits that can be obtained from rigorous piloting of valuation studies. These include achieving optimal experimental designs, but can go beyond this. Here, the piloting was also shown to have influenced the final design of the instrument to be valued, and to have informed how versions of the measure with slightly different wording could be used in the future with the same value set. The work suggests that the same sort of attention should be given to assessment of feasibility and piloting in valuation tasks as it is currently given in randomised trials of complex interventions.

Acknowledgments This work was supported by the European Research Council (261098EconEndLife). The work was conducted whilst Joanna Coast was based at the University of Birmingham. We thank the members of the EconEndLife advisory group and all research participants. We also thank Raymond Oppong and participants at the International Choice Modelling Conference, Sydney, July 2013 and the iHEA congress, Sydney, 2013 for comments on earlier versions of the paper.

Author contributions All authors contributed to the design of the study, interpretation of data and final approval. JC additionally conceived the overall study and wrote the first draft of the paper. TNF additionally conceived the experimental design and analysis plan for the DCE/BWS, acquired and analysed the data, and contributed to revising the paper. EH additionally acquired and analysed the data and contributed to revising the paper. PK additionally oversaw the ethical aspects of the study and contributed to revising the paper.

Compliance with Ethical Standards

EH, TNF and PK declare that they have no conflict of interest. JC received a grant for the work from the European Research Council (261098 EconEndLife).

References

1. Medical Research Council. A framework for development and evaluation of RCTs for complex interventions to improve health. London: Medical Research Council; 2000.
2. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P. Framework for design and evaluation of complex interventions to improve health. *BMJ*. 2000;321:694–6.
3. Craig P, Dieppe P, Macintyre S, Mitchie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ*. 2008;337:a1655.
4. Coast J, Al-Janabi H, Sutton E, Horrocks S, Vosper J, Swancutt D, Flynn TN. Using qualitative methods for attribute development for discrete choice experiments. *Health Econ*. 2012;21(6):730–41.
5. Coast J, Horrocks S. Developing attributes and levels for discrete choice experiments using qualitative methods. *J Health Serv Res Pol*. 2007;12:25–30.
6. Normand C. Measuring outcomes in palliative care: limitations of QALYs and the road to PaLYs. *J Pain Symptom Manage*. 2009;38:27–31.
7. Gomes B, Harding R, Foley KM, Higginson I. Optimal approaches to the health economics of palliative care: report of an international think tank. *J Pain Symptom Manage*. 2009;38:4–10.
8. Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, McIntyre L, Tulsky JA. Factors considered important at the end of life by patients, family, physicians, and other care providers. *J Am Med Assoc*. 2000;284:2476–82.
9. Sutton E, Coast J. Development of a supportive care measure for economic evaluation of end-of-life care, using qualitative methods. *Palliat Med*. 2014;28:151–7.
10. Sutton E, Coast J. ‘Choice is a small word with a huge meaning’: autonomy and decision making at the end of life. *Policy Politics*. 2012;40(2):211–26.
11. Flynn TN, Louviere J, Peters T, Coast J. Best-worst scaling: what it can do for health care research and how to do it. *J Health Econ*. 2007;26:171–89.
12. Louviere JJ, Hensher DA, Swait JD. Stated choice methods: analysis and application. Cambridge: Cambridge University Press; 2000.
13. Flynn TN. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. *Exp Rev Pharmacoecon Outcomes Res*. 2010;10(3):259–67.
14. Brouwer WBF, Culyer AJ, van Exel NJA, Rutten FFH. Welfarism vs. extra-welfarism. *J Health Econ*. 2008;27:325–38.
15. Coast J, Smith RD, Lorgelly P. Welfarism, extra-welfarism and capability: the spread of ideas in health economics. *Social Sci Med*. 2008;67:1190–8.
16. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. Process and methods guide. London: NICE; 2013.
17. Coast J. Strategies for the economic evaluation of end-of-life care: making a case for the capability approach. *Exp Rev Pharmacoecon Outcomes Res*. 2014;14(4):473–82.
18. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-D (EQ-5D-5L). *Qual Life Res*. 2011;20:1727–36.
19. Al-Janabi H, Flynn TN, Coast J. Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Qual Life Res*. 2012;21:167–76.
20. Flynn TN, Huynh E, Peters TJ, Al-Janabi H, Clemens S, Moody A, Coast J. Scoring the ICECAP-A capability instrument: estimation of a UK general population tariff. *Health Econ*. 2015;24(3):258–69.
21. McFadden D. Conditional logit analysis of qualitative choice behaviour. In: Zarembka P, editor. *Frontiers in econometrics*. New York: Academic Press; 1974. p. 105–42.
22. Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. *J Market Res*. 1993;30:305–14.
23. Flynn TN, Peters TJ, Coast J. Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data. *J Choice Model* 2013;6(1):34–43.
24. Louviere JJ, Flynn TN, Marley AAJ. Best-worst scaling: theory, methods and applications. Cambridge: Cambridge University Press; 2015.