Can we value 'capability'? Findings from a pilot study applying a Multi-Attribute Value Method.

Philip Kinghorn[1], Angela Robinson[1], Richard Smith[2]
[1]University of East Anglia
[2]London School of Hygiene and Tropical Medicine

Abstract:
Interest in, and the application of, Sen's 'capability approach' in health economics is increasing. However, a fundamental problem that has been highlighted is whether and how the results of a capability instrument may be valued for use in economic evaluation. The aim of this study was to pilot a Multi-Attribute Value Method for this.

The capability questionnaire contains 42 questions, each relating to a specific functioning; these 42 functionings are grouped under 9 capabilities. The levels of functioning within each question were valued by members of the public, using a simple rating scale. Participants were also presented with a swing-weighting exercise in order to derive weights for the capabilities. Values for the levels of the functionings and weights for the capabilities are combined to arrive at the final scoring system. Discussion was also undertaken during the exercise to explore participants' understanding of the task and the thinking behind their responses. A functional form for the MAV model was chosen based on the qualitative responses of participants.

This paper discusses the issues identified in this study – such as the possible presence of part-whole bias, and the appropriateness of different methods of calculating importance weights – which are likely to have wider relevance for similar studies. An extension to our study suggests that with a prior awareness of these issues, simple steps can be taken to minimise their effect on results.

1. Introduction:

Interest in, and the application of, Sen's 'capability approach' in health economics is increasing [1-7]. However, a fundamental problem that has been highlighted is whether and how the results of a capability instrument may be valued for use in economic evaluation [8]. Within the wider development economics literature capabilities are not valued in the sense that health economists would understand – there is no weighting or numerical value attached. Within health economics, the ICECAP instrument, developed by Coast and colleagues, has a valuation system developed using Discrete Choice Experiment, although this instrument was developed 'backwards' [9]. That is, it arose from a project to apply DCE to aspects of health important to older people, which was then found to be consistent with the Capability Approach. Other work ongoing by Lorgelly and colleagues in Glasgow has yet to determine a valuation technique for their Capability instrument. We have outlined elsewhere a project which

was aimed at developing a Capability instrument for those with chronic pain [10]. In this paper we turn to considering the specific issue of *how* to value the respective capabilities from that instrument.

2. Background:

First, qualitative work during earlier phases of the study produced a list of capabilities of importance to patients with chronic pain. Nine capabilities were identified:

1. To be able to have Self-Respect (SR)
2. To have the opportunity to enjoy social interaction (SI)
3. To be able to fulfil the role of parent and grandparent (PG)
4. Being able to remain physically & mentally active (PA)
5. Being able to have a positive and individual identity (ID)
6. Being able to be independent and to have control (IC)
7. Feeling able to participate in a loving relationship (RE)
8. Being able to enjoy good physical & mental well-being (WB)
9. Being able to take enjoyment from life (EN)

Each capability was summarised in one or two words and used to form the section headings of a questionnaire (abbreviations will be useful later in the paper). Within each Capability a number of functionings represent the actual questions. Under each of the functionings are listed four possible levels. For example:

Figure 1: Example Question

6. Independence & Control

A. Over the past month, I have been able to care for myself (dress, shower, and use the toilet):

- ❑ With no help at all from others and no difficulty
- ❑ With no help at all, but with some difficulty
- ❑ I have needed some help and/or have great difficulty dressing, showering or using the toilet
- ❑ I am completely dependent on others to dress, shower, or use the toilet

In total there are 42 functionings, each with four levels, ranging from no restriction to severe or complete restriction/inability to perform some task or role of importance.

3. Possible Techniques for Weighting Capabilities:

It is suggested within the capability literature that the relative importance of a set of capabilities is determined; it is less clear which valuation techniques should be used or avoided.

Whilst value elicitation techniques, such as SG and TTO, offer one possible means of valuing all states directly, it was felt that their use would be a departure from the theoretical foundations of the Capability Approach. While in health economics, choice-based techniques are thought to be theoretically superior, Sen has been critical of the interpretation of utility in terms of choice [11] and here a choice*less* technique may fit better with the theory on capability. This – and the apparent simplicity of the method – led us to explore the use of a specific Multi-Attribute Value Method, first developed by Peacock *et al.* [12].

4. The Multi-Attribute Value Method

Peacock et al. [12] developed a multi-attribute value method in the context of assessing mental health services in South Australia. While this context is clearly different to ours, the methods used by Peacock *et al.* are relevant in a wide range of settings and we outline the methodological steps below:

1. Identifying attributes in the MAV function
2. Describing attributes
3. Scaling attribute levels
4. Quantifying trade-offs between attributes
5. Evaluating Programmes
6. Combining Attribute Scores

In the study by Peacock *et al.* steps one and two were completed by an advisory panel. Step three involved the panel scaling the different levels within the attributes, with their relative importance being determined on a 0-100 scale. 'Attribute worst' and 'attribute best' levels were placed on the rating scale at 0 and 100 respectively to define the endpoints of the measurement scale for each attribute. Panel members then placed the remaining/intermediary levels on the scale between these end points. The average of panel members' responses for each intermediate level was calculated and used to calibrate the scale for each attribute [12].

The panel then assessed the relative importance of each attribute, through the use of a 'swing-weights' method (step 4). In this approach, respondents are asked how much an attribute contributes to overall utility relative to other attributes by comparing

hypothetical programmes that swing between the worst and best levels in each attribute. They then estimate the change in utility that would result from changing each attribute from its worst to its best level using a rating scale with endpoints of 100 (all at best) and 0 (all at worst) [12]. Both variations of the Swing-Weights method were used. With top-down, the starting point is 'all at best' and the loss associated with some attributes deteriorating to at worst is calculated; with bottom-up, the starting point would be all at worst and the gain associated with attributes improving to at best is calculated.

Peacock *et al.* use a relatively straightforward method for calculating importance weights for their attributes:

1. Where a scenario represents an improvement (deterioration) in one attribute individually, then the mean gain (loss) associated with that one scenario is taken as the mean utility gain (loss) associated with that particular attribute.
2. Where there is a two-step process (i.e. where scenario A represents a change in the first attribute individually and scenario B represents a change in the first attribute and a second together) then the mean gain (loss) associated with the second attribute is found by calculating the difference between the second shift and the first. For bottom-up, the gain associated with the second attribute improving to at best would be calculated using the formula:
   Mean Gain = Score for B – Score for A.
3. The Importance weight for an attribute is simply the mean of the mean gain from bottom-up and the mean loss from top-down. Importance weights for attributes are calibrated so that they sum to one hundred.

Next, the different programmes were evaluated according to how well they performed with respect to each attribute. Finally, panel scores for each attribute were combined using the MAV model to calculate the overall benefit score for the programme. The panel decided upon a functional form for the MAV model that would allow Individual Health being at worst to reduce the overall score for a programme to zero. The functional form is a combination of additive and multiplicative elements. The formula (including the calibration) is:

$$B=(H/100)*[M + ((100 - M)/\Sigma W_i)(W_c D_c/100) + ((100 - M)/\Sigma W_i)(W_e D_e/100)]$$

Where B is the overall benefit score, D_c and D_e are scores for Community Health and Equity on their respective measurement scales, W_c and W_e are the importance weights for the attributes Community Health and Equity, H is the score for Individual Health on its measurement scale and M is the importance weight for Individual Health.

5. Adapting the MAV Method to Score the Capability-Based Questionnaire

Overview:

In the case of our study, the 'panel' was members of the general public who came to the university and were allocated to one of six groups. Groups lasted for approximately 90 minutes and were run in three sections: participants were welcomed and given the background to the study; participants were given the scaling exercise (step 3), which was explained to the group as a whole, completed by the participants individually and then discussed within the group; participants were given the swing-weighting exercise (step 4), which was explained, completed by them individually and then discussed within the group.

Sample Size:

Step three involved participants scaling the different levels within the functionings. There are 42 functionings within the questionnaire, each with 4 levels. Each of the levels within each of the 42 functionings had to be valued directly. It seemed reasonable to expect respondents to consider a maximum of seven functionings, which meant that a minimum of six groups was required. It was decided that 5 to 6 participants would be recruited for each group, giving a target sample size of between 30 and 36 participants.

Recruitment:

Most participants were recruited from a database of research volunteers containing the details of members of the public living in the Norwich area who had previously indicated that they would be willing to participate in academic research. A small number of undergraduate and postgraduate students were also recruited for the study.

Scaling Exercise for Levels:

Participants were initially presented with an example functioning/question and an example scale, which were introduced and discussed by the researcher (see Figure 1 for

an example question). It was explained that the top and bottom levels formed the endpoints for the scale and that two intermediary levels fitted on the scale at points between these extremes.

The researcher then presented a scenario completely unrelated to chronic pain in order to further familiarise participants with the rating scale. A number of points were stressed as the example was discussed, namely that the respondent should answer in a way that reflected their *own personal* preference; that the respondent should answer as if *they* were the one's facing the scenarios described; that it did not matter in what order the intermediate levels were placed on the scale.

Respondents then worked through seven questions, taken from the forty-two contained on the questionnaire. For each question they were given a description of the two intermediate levels (labelled A and B) and a rating scale of 0 to 100. On each scale the best level was pre-printed at 100 and the worst level was pre-printed at 0. Every participant within a group was given the same seven questions.

Once every participant had completed the task they were encouraged to discuss their answers, and a series of open-ended questions were used to initiate this discussion. This allowed us to establish whether respondents had understood the questions and gain some insights in to why they had answered in the way that they had.

Swing-Weighting exercise:

The purpose of the swing-weighting exercise was to quantify trade-offs between the nine capabilities. A particular capability is said to be 'at best' when the top level is selected by the patient on every functioning listed under that capability, and, vice-versa, the capability would be said to be 'at worst' when the worst-level for every functioning listed under that capability is selected. Participants were only asked to consider scenarios in which each capability was at one of these extremes.

When all nine capabilities are 'at best' then a patient can be said to have the best possible quality of life and, vice-versa, when all nine capabilities are 'at worst' then the respondent should be thought of as having the worst possible quality of life; the evaluative space here is well-being freedom. The researcher stressed to participants that the scenario in which all categories are 'at best' is fixed at 100, and that the

scenario in which all categories are 'at worst' is fixed at zero on the scale. Participants were asked to consider four scenarios, the first (starting) scenario was either 'all at best' (at 100) for top-down or 'all at worst' (at zero) for bottom-up. Participants were read a description of the nine categories in their opening state and encouraged to seek clarification if the descriptions were at all unclear. The following three scenarios involved some categories deviating from this starting position, and participants decided where on the scale these three scenarios should be placed; the task was completed by participants individually.

A maximum of three capabilities were allowed to 'swing' (between best and worst) in any single scenario, the idea being that if the remaining six stayed constant the task would be less cognitively demanding.

We initially assumed that there is mutual independence between the nine capabilities. If this is the case, then preferences for varying degrees of achievement on any one capability can be assessed after fixing achievement on all other capabilities at any convenient level [13]. In turn, it was also, at first, assumed that the appropriate functional form for the model would be additive.

6. Results

Table1: Summary of Characteristics of Participants by Group:

Group	No. Male Participants	No. Female Participants	Min Age	Max Age	Mean Age
1 (TD)	3	2	49	65	56
2 (BU)	0	2	30	66	48
3 (TD)	3	2	34	67	57
4 (BU)	2	4	25	64	45
5 (TD)	3	2	28	63	50
6 (BU)	0	4	23	41	29
Total	11	16	23	67	48

Characteristics of the 27 participants who took part in the study are reported by group in Table 1. TD denotes that a group was presented with the top-down approach; BU with the bottom-up approach.

Scaling of the Levels within Functionings:

Scores for a *selection* of levels are presented in Table 2. If we take the example from Figure 1 as an illustration, then we can see that the top level, being able to care for

one's self, is fixed at 100. The second level (level A), being able to care for one's self with no help at all, but with some difficulty, was given a mean score of 75. The third level (level B), needing some help and/or having great difficulty caring for one's self, was given a mean score of 20. The bottom level, being totally dependent on others, is fixed at zero.

Table 2: Scores for levels within Functionings:

Functioning:	Level A			Level B		
	Range	Mean	Median	Range	Mean	Median
6A Self-Care	10	75.0	75.0	0	20.0	20.0
6C Walking	40	70.0	70.0	40	40.0	50.0
6E Stairs	40	56.0	70.0	40	26.0	30.0
6G Cook & look after the home	50	74.0	80.0	40	40.0	40.0
8A Pain	40	75.6	80.0	18	13.6	10.0
8B Effects of medication on future health	20	70.0	70.0	30	26.0	30.0
8C Side-effects of medication	10	75.0	75.0	0	30.0	30.0
8D Sleep	10	73.3	70.0	15	29.2	30.0
8E Physical exercise	30	76.0	80.0	20	34.0	40.0
8F Depression	20	67.8	68.5	10	23.3	20.0

What is noticeable is that in many cases there is some significant degree of polarization between the scores for A and those for B. From the qualitative data, it seems that a likely cause of this is the descriptions of the levels, as the description of A tends to be quite positive ("I *can do* X with some difficulty") whereas the descriptions of B tend to be quite negative ("I can do *little* of X", or "I *cannot* do X at all some days). Participants felt that, although there may be 'some' difficulty doing something of importance, it was positive that the outcome was still achievable, in this way the restriction tended to be viewed as an inconvenience or a frustration. Participants tended to view a move to level B as more of a loss or as a much more significant restriction in freedom.

Swing Weighting:

In terms of qualitative observations, a problem was encountered where participants found that when certain capabilities were swung to the extremes of 'at best' or 'at worst' there was some contradiction in terms of what an individual was described as being able to do in some categories and what they were unable to do in others. For example, with Well-being at worst, the individual would have constant pain, difficulty sleeping most nights, be unable to exercise physically and have severe depression. And

yet with other capabilities at best they would be described as being able to pursue hobbies and interests, take enjoyment from life and not worry about future health. Although participants were still able to complete the task in a way that they felt was meaningful, some scenarios were changed after the first top-down and the first bottom-up group in an attempt to make them more plausible. We return to this issue in the discussion.

Summaries of the Quantitative Data:

The gains measured using the bottom-up approach are presented in Figure 3, with scenarios in which single and multiple capabilities have improved to at best, from a starting point of all being at worst. The capabilities which are at best are listed across the x-axis, in each case all other capabilities (i.e. those which are not named) will be at worst.

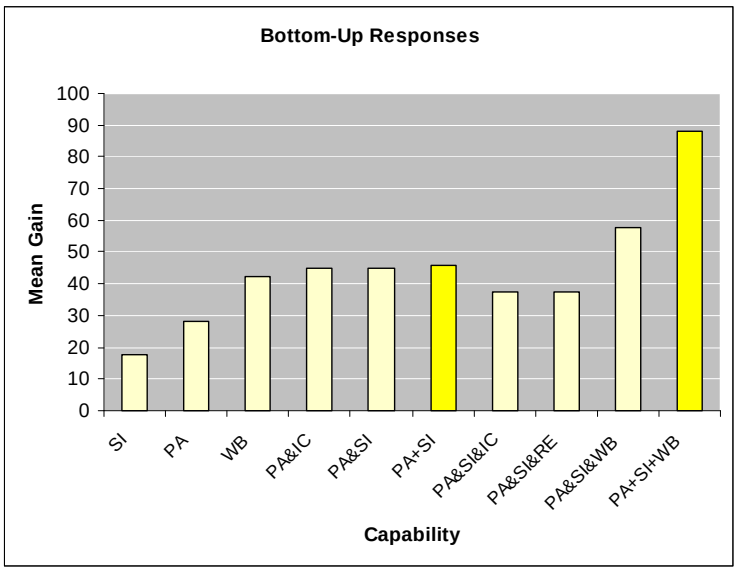

Figure 3: Mean Gains Obtained Using Bottom-up

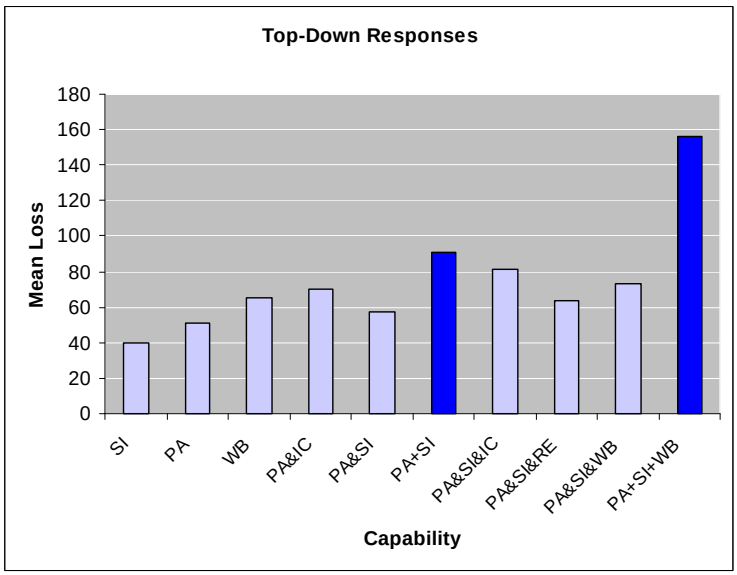

Figure 4: Mean Losses Obtained Using Top-down

Whereas, for example 'PA & SI' is the score for the scenario in which PA and SI were jointly at best (for bottom up), 'PA + SI' (the darker coloured bar) is the *sum* of the scores for the separate scenarios in which PA and SI were individually at best.

Mean losses associated with the top-down approach are presented in Figure 4. Mean losses are calculated by subtracting the mean score for a scenario from 100.

It may be useful to set out some assumptions about what we would expect to see within the results. We will work through these assumptions in turn.

Relative Importance:
First, we have assumed that the capabilities will not all have equal importance, and so we would expect to see the following relationship:

1. $SI \neq PA \neq WB$

Figure 3 demonstrates that capabilities are viewed by respondents as having different degrees of importance, with the score associated with WB being over twice that for SI. Again, in Figure 4, if we look at the capabilities SI, PA and WB, WB appears to be the most important capability of the three, and PA the second most important.

Dominance:
We would expect participants to prefer scenarios where two or more capabilities are at best to scenarios in which only one of those capabilities was at best individually, or at least to be indifferent between the two. In some respects the mean gains expressed in Figure 3 appear to be as expected; so, for example the scenario in which both PA and SI are at best (PA&SI) has been placed at a higher point on the scale than the separate scenarios in which PA and SI are individually at best. Other aspects of Figure 3, however, are not as expected. The mean score for the scenario in which PA and IC were jointly at best and all other capabilities at worst was 45; but, if we consider the scenario in which PA, IC and WB are at best and all others are at worst, then this apparently better scenario has a lesser mean score of 37.5. The scenario in which PA, SI and RE are at best is also given a mean score of 37.5, which again is less than the mean score for the scenario in which only PA and SI are best (45). All tests of dominance are met within Figure 4.

We can also consider individual responses. Participants considered four scenarios, the first scenario (A) was fixed on the scale at zero (or 100), depending on whether they were completing bottom-up or top down; in all cases scenario B involved only one capability changing to at worst (at best), scenario C involved three capabilities changing, and in scenario D two capabilities were changed to be at worst (best). We would therefore expect to see the following responses for bottom-up (and top-down):

2. $C \geq B$ ($B \geq C$)
3. $C \geq D$ ($D \geq C$)

Scenarios B and D cannot directly be compared as the capabilities being considered were not the same, e.g. Scenario B: SI; Scenario C: SI, IC, PA; Scenario D: IC, PA.

Table 3 summarises how many participants gave responses which passed the above tests of dominance, and it can be seen that the overwhelming majority of responses do pass the dominance tests.

Table 3: Individual Dominance Tests:

Test	Number of Passes	Percentage Passed
Bottom-Up		
$C \geq B$	12/12	100
$C \geq D$	11/12	91.7
Top-Down		
$B \geq C$	14/15	93.3
$D \geq C$	13/15	86.7

Now consider the results at a group level. The scenarios presented to the groups, and the mean responses from the groups are presented in Table 4.

Clearly, for top-down the capabilities listed in the scenarios would have been at worst, while all others were at best, and vice-versa for bottom-up. As discussed, we would expect a scenario in which a particular capability is at worst individually to be preferred to one in which that capability and others are jointly at worst. There is only one group (Group 2) where this was not the case when we look at the mean of the responses, but this is due to the small size of the group and the influence of one participant. Responses failing the tests of dominance outlined above were then removed; the mean of the remaining responses are presented in Table 5.

Table 4: Mean Scores by Group:

Group	Scenario B	Score (B)	Scenario C	Score (C)	Scenario D	Score (D)
1 (TD)	SI	54.0	SI, IC, PA	**19.0**	IC, PA	**30.0**
2 (BU)	SI	12.5	SI, IC, PA	**37.5**	IC, PA	**45.0**
3 (TD)	WB	**35.0**	WB, SI, PA	**27.0**	SI, PA	43.0
4 (BU)	WB	**38.8**	WB, SI, PA	**55.0**	SI, PA	41.3
5 (TD)	SI	66.0	RE, SI, PA	**36.0**	PA	49
6 (BU)	SI	50.0	RE, SI, PA	**60.0**	PA	42.5

Table 5: Mean Scores by Group, with 'cleaned data':

Group	Scenario B	Score (B)	Scenario C	Score (C)	Scenario D	Score (D)
1 (TD)	SI	54.0	SI, IC, PA	**19.0**	IC, PA	30.0
2 (BU)	SI	5.0	SI, IC, PA	**25.0**	IC, PA	15.0
3 (TD)	WB	**40.0**	WB, SI, PA	**26.3**	SI, PA	47.5
4 (BU)	WB	**46.7**	WB, SI, PA	**66.7**	SI, PA	46.7
5 (TD)	SI	62.5	RE, SI, PA	**26.3**	PA	47.5
6 (BU)	SI	50.0	RE, SI, PA	**60.0**	PA	42.5

Additivity:

Given our assumption that the model will be additive, we would expect to see the following trends in the data:

4. PA & SI = PA + SI
5. PA & SI & WB = PA + SI + WB

If we sum the independent gains associated with PA and SI (PA+SI), as in Figure 3, we see that the resulting gain is very close to the score given to the scenario in which both PA and SI are at best together (PA&SI); 45.8 vs. 45. The addition of WB increases the mean score yet further, although the mean score for this scenario is significantly less than the sum of the mean scores for the three scenarios in which PA, SI and WB are individually at best, indicating some diminishing gains.

Although mean responses seem to follow an expected pattern in the one case outlined above, when we have three capabilities changing in a single scenario we see that there are significant diminishing gains. In Figure 4 there appears to be significant diminishing losses associated with the scenarios in which two or more capabilities swing to at worst. If we consider the results at a group level (Tables 4 and 5), then, even in our 'cleaned' data, where Well-Being was entered into a scenario it seems to have resulted in the bottom-up scenarios being placed higher on the scale than top-down scenarios, in which less capabilities were 'at worst'. It can also be seen that one

individual capability changing to at worst in top-down leads to a fall approximately half way down the scale (more so for Well-being), a huge drop considering eight other capabilities are still at best. What is clear is that the shifts up (or down) the scale that are associated with apparently minor changes from the starting point are so large that respondents are leaving only a very small distance on the scale for further shifts associated with further and potentially much more significant changes.

It can also be seen that there is a significant 'overlap' between the distance moved down the scale when three capabilities change to at worst, and the distance moved up the scale when three capabilities change to at best. Even despite any differences we may expect due to reference point bias, the pattern in Tables 4 and 5 appear to suggest both that respondents over-react to minor changes and, as discussed earlier, that further changes then have diminishing effects. Although it may be impossible to discover through this study alone why changing just one capability proved to have such a large impact; four possible explanations are listed below:

Issues relating to the bottom-up method:

1. The starting position (i.e. all capabilities at worst) is deemed to be so awful that any improvement will be very significant, and so respondents place scenarios in which there is only a modest improvement at a high position.
2. If one capability is at best then respondents find it implausible that others could simultaneously all be so bad, so they adjust everything else upward.

Issues relating to top-down:

3. Well-being is itself simply a hugely important capability, which people expected would affect other aspects of life. One respondent thought that Well-being alone being at worst would be like a 'dark cloud' that would cast a shadow over every other aspect of life, another spoke of well-being deteriorating as triggering a 'domino effect'.

Bottom-up and top-down:

4. Respondents were only asked to consider changes in up to three of the nine capabilities, therefore they perhaps used the full length of the scale without necessarily considering the entire scenario; they are not thinking about the fact that they have not left room on the scale for scenarios in which more than three

capabilities change. Respondents are not compensating enough for the fact that 6, 7 or 8 capabilities remain exactly as they were before.

It appears that some of what we are seeing in the results may be a form of misspecification (part-whole) bias, usually seen in contingent valuation studies, and thought to occur when: "respondents are unable to differentiate between benefit subcomponents or between the subcomponents and the value for all types of benefits" [14, p251]. In short, the respondent confuses some broader or wider reaching entity with the entity that the researcher wants to value. What is seen, therefore, is that respondents state much the same preference for different scales of projects/benefit; this phenomenon has also been termed 'embedding'.

7. Calculating Importance Weights:

The 'mean response method' used by Peacock *et al.* to calculate importance weights is outlined on page 6 above. We have three capabilities (PA, SI and WB) for which we can simply apply step one of their method, so for example, when only Social Interaction was at best using bottom-up, the mean gain was 17.5, the mean loss when only Social Interaction was at worst in the top-down method was 40, so the importance weight for Social Interaction would be 28.75. The weights for these three capabilities are presented in Table 6.

Table 6 Weights for Selected Capabilities by Mean Responses Method

Capability	Bottom up method	Top down method	Weight
Social Interaction (SI)	17.5	40	28.75
Physically & Mentally Active (PA)	28.3	51	39.65
Physical & Mental Wellbeing (WB)	42.5	65	53.75
Independence & Control (IC)	16.7	19	17.85

Also contained in Table 6 is an importance weight for Independence & Control (IC), which has been calculated using the process, outlined in step 2 on page 6. So, for top-down, the loss associated with PA alone was 51, whereas the loss associated with PA and IC was 70, so for top-down the mean loss for IC was 19 and mean gain 16.7. The importance weight for IC is therefore 17.85.

What is interesting is that, despite IC being reported as being an important capability by respondents in the qualitative work, the importance weight for IC is the lowest of the

four and quite a lot smaller than that for SI. It was noted earlier that there appeared to be diminishing gains (losses) for scenarios in which increasing numbers of capabilities are at best (worst), and so it seems inevitable that we will get lower importance weights calculated according to step 2 than we will by using step 1 and therefore importance weights calculated using these two different steps in the mean response method do not appear to be comparable.

In order to explore this idea further let us calculate the importance weight for SI through step 2. When PA is individually at best the mean gain is 28.3, and when PA is individually at worst the mean loss is 51. When PA and SI are both at best the mean gain was 45, so the mean gain associated with SI, calculated according to step 2, must be 16.7. When PA and SI are both at worst the mean loss was 57, so the top-down mean loss associated with SI must be 6. So, the overall importance weight for SI must be 11.35, which is significantly less than the importance weight calculated for the same capability using step 1 (28.75), and would rank SI below PA.

Next, we can explore the outcome of using step 2 with swings involving more than two capabilities; we will calculate the importance weight of IC. With bottom-up, we get a mean gain associated with IC of -7.5. From top-down we get a mean loss associated with IC of 24. Here then, the importance weight for IC would be 8.25.

It is clear, therefore, that we should have serious concerns about using what has here been termed the Mean Response Method, at least when we have relatively large numbers of capabilities (attributes). It would seem sensible to compare importance weights which have all been calculated from 'individual swings' (i.e. through steps 1 and 3 only) as this way we can be sure that the weights are all comparable. But, if we were only to use gains (losses) associated with scenarios in which just one capability improves (deteriorates) then it would seem that we would only be capturing any interaction or form of trade-off between the capabilities in an extremely limited way.

A more appropriate method of calculating the importance weights may be to use regression analysis, with the coefficients for the capabilities being interpreted as importance weights. Table 7 provides a comparison of results calculated using the two different methods. Numbers in brackets represent rankings.

Table 7: Comparison of Importance Weights for Capabilities by Mean Responses and by Regression Analysis

Capability	Weight from Mean Response	Weight from Regression Analysis
Social Interaction (SI)	28.75 (3)	13.12 (4)
Physically & Mentally Active (PA)	39.65 (2)	25.39 (2)
Physical & Mental Wellbeing (WB)	53.75 (1)	28.16 (1)
Independence & Control (IC)	17.85 (4)	16.13 (3)

Only a small number of scenarios were presented to the groups, as the objective was to pilot the methods, rather than embark on a full data collection exercise. Because of this, we are not able to calculate importance weights for all nine capabilities through regression. Instead, we have 'filled in the gaps' by placing the remaining capabilities on a scale, given the relative position of those we were able to calculate from the data. Raw weights are therefore able to be presented for all nine capabilities in Table 8. Values in bold are those calculated directly from the data.

Table 8: Calibrated Importance Weights

Capability	'Raw' weight	Calibrated Weight
Self-Respect (SR)	13.00	7.74
Social Interaction (SI)	13.12	**7.82**
Role of Parent/ Grandparent (PG)	14.00	8.34
Physically & Mentally Active (PA)	25.39	**15.12**
Identity (ID)	12.00	7.14
Independence & Control (IC)	16.13	**9.61**
Relationships (RE)	15.00	8.93
Physical & Mental Wellbeing (WB)	28.16	**28.16**
Enjoyment (EN)	12.00	7.14

A first step in calculating an overall capability score (CS) for a patient is to weight the responses (entered as scores – see Table 2) using the calibrated importance weights for the corresponding capability. CS is then calculated by taking an average of the weighted scores across the Functionings corresponding to an individual Capability, i: call this average DS_i. We also need to take an average of the scores for those Functionings corresponding to Well-being: call this WB. The formula for calculating CS is then:

$$CS=(WB/100)*(W_{WB} + DS_{SR} + DS_{SI} + DS_{PG} + DS_{PA} + DS_{ID} + DS_{IC} + DS_{RE} + DS_{WB} + DS_{EN})$$

Using this formula, the best possible Capability Score is 100, and the worst is zero. Using our calibrated importance weights, and the importance values for the levels with Functionings, the Capability Score for a patient ticking level A for every question on

the Questionnaire would be 63.814, and the Capability Score for a patient ticking level B on every question would be 13.316. Notice that the large gap here is due to the way participants in the groups rated the levels within the Functionings, where level B was generally thought to be much worse than level A. Limitations in terms of the actual data are clear.

8. Exploring Methods to Minimise Part-Whole Bias

A number of possible weaknesses were identified with the original study design. Respondents completing the Swing-Weights task were given a full description of all of the capabilities at their starting level (i.e. Scenario A) and were then presented with a new description of only those capabilities that changed in the following 3 scenarios. First, it was felt that had respondents completing top-down been given a full description of the worst possible scenario as part of the initial explanation they would have appreciated the huge contrast, and just how serious and undesirable this scenario at the bottom of the scale would be. This may have impressed on respondents that in the broader scheme of things the scenarios they were considering were relatively 'minor' deviations from the best scenario. Second, there was a concern that respondents were focussing too heavily on the descriptions of the changed capabilities and not fully taking into consideration that many of the other capabilities were still at their best level.

It was felt, and was justified earlier, that scenarios in which more than three capabilities changed simultaneously would become complex, and that limiting the number of capabilities changing in any one scenario would make the task less cognitively demanding. Also, the scenarios were presented in such a way that the first scenario involved only one capability changing, the second involved three that had changed, the third only two. Given the results, it was hypothesised that were respondents in future groups given an initial scenario in which there was a more dramatic change then this would help them to 'get a feel' for the full scope of the scale. From such a dramatic initial change, the scenarios which previous respondents had considered would surely appear less significant.

Given a suspicion that what we are seeing in the results so far is a form of part-whole bias, and a reflection on the original study design, the study was extended and a number of changes made to the design and delivery of the Swing-Weighting Exercise:

1. Participants were only asked to complete the Swing-Weighting Exercise (but given the same introductions and explanations as before).
2. Only the top-down method was used.
3. Participants were given a verbal description of the worst possible scenario before they started. It was made clear that this scenario was fixed at zero.
4. Participants were presented with a full written description of every category within the scenario (those at best as well as those at worst).
5. Four scenarios (five including A which was fixed on the scale at 100) were used this time instead of three (four including A).
6. In the first scenario that respondents had to place on the scale (scenario B) 5 categories had changed to "at worst". In the initial scenarios a maximum of only three scenarios were allowed to change.

Some changes, such as only asking respondents to complete the swing-weighting task and only using top-down, were made for convenience, to reduce the workload and time needed to run additional groups. The difference in results is quite striking; when only one capability is at worst, then the mean loss was nearer to one quarter (24.4) in the study extension, compared to approximately one half (49) previously. When two capabilities were at worst in the study extension the mean loss was approximately one third (32.2), compared to nearer two-thirds (63.5) previously.

9. Discussion

A multi-attribute value method has been used to scale the levels within functionings and quantify trade-offs between capabilities taken from a capability-based questionnaire, developed to assess quality of life in patients with chronic pain. We found that:

1. Results from a swing weighting exercise suggest that a purely additive functional form for the MAV model would be inappropriate; it appears that we cannot assume independence between the capability 'Well-being' and the other eight capabilities. It was therefore decided to adopt the functional form used by Peacock *et al.* where well-being at worst can reduce the capability score to zero, but there is an additive relationship between the remaining eight capabilities.
2. Part-whole bias was present in the results from the swing-weighting exercise.

3. Importance weights calculated using steps two and three of the 'mean response method' were found to be markedly smaller than those calculated through steps one and three of that method. When importance weights were instead calculated using regression analysis they appeared to be more accurate, and the ranking of capabilities fitted with that reported by participants in the discussion.

We would judge the MAV model to be useful and relatively straightforward to implement, and for respondents to complete. Where problems were encountered they were no more severe than would be expected in most pilot studies, and there is good indication that steps can be taken to resolve them before any wider study might be launched. One such 'issue' was the likely presence of Part-whole bias, so far defined and discussed as a phenomena present within contingent valuation studies [15]. If readers agree that it was part-whole bias which was present in the original results then one of the most prominent objections to the use of CVM [16] will look less valid, and other valuation techniques may be just as susceptible to this weakness.

An issue that will require consideration whenever multi-attribute health state systems are valued - particularly those with a sufficiently large number of possible states so as to prohibit the valuation of each directly - is that of independence between the attributes. It was found in this pilot study that the capability 'Wellbeing' cannot be assumed to be mutually independent from the other eight attributes, and, in reality, still further interaction may be present between the remaining capabilities. While we were able to adopt a functional form for the MAV Model which appears to allow appropriate interaction, the apparent lack of independence caused difficulties during the Swing-Weighting exercise. Furthermore, had we not included a qualitative element to the study, the interaction would not have been so clearly evident.

Additive functions which have been used in the quality of life literature include the Quality of Well-being Scale, the EQ-5D, the SF-6D and the 15D Health-Related Quality of Life Instrument [12]. So, for example, in EQ-5D, extreme pain and discomfort and/or extreme anxiety and depression could not independently impact catastrophically on overall quality of life.

We face something of a trade-off between the selection of a functional form that is appropriate given our underlying assumptions, and the use of a simpler and more

practical form which may be deemed 'adequate'. Keeney and Raiffa [13] advise that the additive and multiplicative functions appear to be the more practical for models involving four or more attributes. "Even when the requisite assumptions do not precisely hold over the domains of all the attributes, it may be a good approximation to assume they do, or it may be reasonable to integrate different additive and multiplicative utility functions...". This clearly fits with our handling of the Capability 'Wellbeing' in the MAV Model.

References

1. Cookson, R., *QALYs, and the capability approach.* HEALTH ECONOMICS, 2005. **14**(8): p. 817-829.
2. Verkerk, M.A., J.J.V. Busschbach, and E.D. Karssing, *Health-related quality of life research and the capability approach of Amartya Sen.* QUALITY OF LIFE RESEARCH, 2001. **10**(1): p. 49-55.
3. Lorgelly, P., et al., *The Capability Approach: developing an instrument for evaluating public health interventions*. 2008, University of Glasgow.
4. Grewal, I., et al., *Developing attributes for a generic quality of life measure for older people: Preferences or Capabilities?* Social Science & Medicine, 2006. **62**: p. 1891-1901.
5. Coast, J., et al., *Developing an Index of Capability for Older People: A New Form of Measure for Public Health Interventions?,* in *Future of Health Conference*. 2006: Cambridge.
6. Coast, J., R.D. Smith, and P. Lorgelly, *Should the capability approach be applied in health economics?* HEALTH ECONOMICS, 2008a. **17**: p. 667-670.
7. Coast, J., R.D. Smith, and P. Lorgelly, *The influence of capabilities on health care decision making in the UK.* Social Science & Medicine, 2008c. **67**: p. 1190-1198.
8. Coast, J., et al., *Valuing the ICECAP capability index for older people.* Social Science & Medicine, 2008. **67**: p. 874-882.
9. Coast, J., et al., *Investigating Choice Experiments for Preferences of Older People (ICEPOP): evaluating spaces in health economics.* Journal of Health Services Research Policy, 2008b. **13**(Suppl 3): p. 31 - 37.
10. Kinghorn, P. and R. Smith, *From Theory to Practice: Are we capable of operationalising the Capability Approach*, in *Health Economists' Study Group*. 2008: Norwich.
11. Sen, A., *The Standard of Living*. 1987, Cambridge: Cambridge University Press.
12. Peacock, S.J., et al., *Priority setting in health care using multi-attribute utility theory and programme budgeting and marginal analysis.* 2007.
13. Keeney, R.L. and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. 1993, Cambridge: Cambridge University Press.
14. Mitchell, R.C. and R. Carson, *Using Surveys to Value Public Goods*. 1989, Washington, D. C.: Resources for the Future.
15. Kahneman, D. and J.L. Knetsch, *Valuing Public Goods: The Purchase of Moral Satisfaction.* Journal of Environmental Economics and Management, 1992. **22**: p. 57-70.
16. Svedsater, H., *Contingent Valuation of Global Environmental Resources: Test of perfect and regular embedding.* Journal of Economic Psychology, 2000. **21**: p. 605-623.