Prepared for discussion at HESG Bangor 2011
Work in progress – not for citation or distribution

Completion of self-reported health and capability measures: a think-aloud study

Hareth Al-Janabi*, Thomas Keeley, Paul Mitchell and Joanna Coast

Health Economics Unit, University of Birmingham

Aim: Self-reported outcome data is an important input into economic evaluation, but little is known about individuals' interpretation of, and problems with, outcome questionnaires. As part of the validation process for the ICECAP-A capability wellbeing measure, this study sought to identify the frequency and nature of completion problems that individuals encounter with the ICECAP-A and EQ-5D.

Methods: Thirty-four cognitive (think-aloud) interviews with the general public were carried out to explore the process of completing the ICECAP-A and EQ-5D. Following each think-aloud interview, a semi-structured interview was conducted to explore, in more detail, the nature of any completion problems. Four independent raters coded transcripts for four types of error (comprehension, retrieval, judgement and response) and 'struggles'. Constant comparative analysis is being used to analyse both the think-aloud elements of the interview and the subsequent semi-structured interview.

Results: Six (3.5%) out of the 170 segments of the ICECAP-A were associated with an error and 10 (5.9%) with 'struggle'. Twelve (5.9%) out of the 204 segments of the EQ-5D were associated with an error and 3 (1.5%) with struggle. The most common error occurred when participants articulated a desired response that was not available on the questionnaire.

Conclusions: Both questionnaires appear to be easily understood by the majority of respondents. However, individuals do appear to perceive some problems with response options (for the EQ-5D, especially) and subjectivity (for ICECAP-A, especially). We would be interested in the views of HESG members as to whether these problems are concerning, and if so, how they might be addressed.

---

* corresponding author at Public Health Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, 01214158483, h.al-janabi@bham.ac.uk

1. Background

Self-reported data on an individual's health, quality of life and/or preferences is needed to estimate the benefits of healthcare for economic evaluation. The calculation of benefits in terms of quality adjusted life years (QALYs), for example, requires data on an individual's health status, in order to weight future life years. Individuals (normally patients) typically report their own health, through a short standardised questionnaire such as the EuroQol, SF-36 or Health Utilities Index (1). There is a burgeoning literature that demonstrates that these instruments mostly pass statistical tests of feasibility, reliability and validity (2;3). However, despite the fact that individuals are known to encounter difficulty with standardised questions (4), there has been little evaluation of the meaning of individuals' answers and the degree to which individuals encounter problems in completing the measures.

Unlike in interview-administered questionnaires, in self-completed survey questionnaires there is often no opportunity to clarify the meaning of the question (5). Even seemingly straightforward questions can be interpreted in a number of ways. Problems may become more acute when the survey terms used are relatively vague, which may be a problem in health economic evaluation, where preference-based measures generally attempt to span a broad concept like health in as few items as possible. Other issues that can lead to confusion include double questions embedded within a single question and questions that contain unfamiliar or irrelevant terms.

A second type of problem with survey questions (especially relevant for quality of life questions), is that individuals adapt their lives in response to major events, such as illness (6). As a result of adaptation, the same individuals may answer a given survey question in different ways over time. Known as response shift (7), individuals may change their internal scales, values or conceptualisation of aspects of quality of life in response to changes in their life. This can lead to problems in comparing data collected for an individual over time.

Many of the issues in collecting valid survey data apply whether the data are on health status or wellbeing more broadly defined. Although health related quality of life measures dominate in economic evaluation, measures of broader wellbeing are being proposed (8;9). It is important to establish, not only whether such measures have appropriate statistical measurement properties, but also that the data derived are meaningful. One group of wellbeing measures are the ICECAP capability measures (10;11), developed to provide a broader assessment of the wellbeing impacts of health and social care interventions. The ICECAP-A (for all adults) measure (see appendix A) records an individual's capability to function, as opposed to their functioning *per se*. Briefly the ICECAP-A measure asks individuals to self-report their capability across five dimensions of their life: stability, attachment, autonomy, achievement and enjoyment. The development process for the ICECAP measures involved qualitative work to develop lay terminology for the five items in the measures (10-12). However, the broad nature of the concepts and the 'capability' wording of the questions suggest that establishing the meaning of the terminology is particularly important.

This study sought to identify the frequency and nature of completion problems that individuals encounter with the ICECAP-A capability measure, contrasting these with the widely used EQ-5D health measure (13). It was hypothesised that the broad terms and capability wording of the questions in the ICECAP-A measure may result in higher levels of misunderstanding relative to the EQ-5D.

2. Methods

Think-aloud interviews, with a general population sample of adults, were used to explore the completion of the measures. This section provides a brief overview of the methodology, going on to describe the sampling, conduct and analysis of the interviews.

2.1 Methodology

Think-aloud (or cognitive) interviews have been used to explore the process of task completion, initially in relatively complex, multi-step tasks, such as playing chess or unscrambling an anagram (5), but increasingly in understanding the completion of survey data (14-16). Respondents are asked to verbalise their thoughts while completing a task. The verbal information then provides an insight into the process of completing the task; potentially enabling the identification of the point at which any problems are encountered. Verbalisation should not normally interfere with the task at hand and affects cognitive processes *"only if the instructions require verbalisation of information that would not otherwise be attended to"* (17) p. 215. Verbalisation can either be concurrent (i.e. whilst completing the task) or retrospective (after the task). Concurrent methods have been shown to outperform retrospective methods in terms of both the amount of information generated and the degree of insight into decision-making processes (18). Although think-aloud tasks were not specifically designed for questionnaires, they may be well-suited to examining self-complete instruments (5), and particularly helpful in establishing content validity (19).

2.2 Sampling

Given the potential for the ICECAP measure to be used outside healthcare, individuals were sampled for interview from the general population. Twenty-four individuals were potentially available, as they had indicated a willingness to take part in the ICECAP-A research, but were not used in the interviews to develop the ICECAP-A measure. To expand the sample frame, an invite and simple screening questionnaire were sent to 600 randomly selected[1] individuals from four electoral wards in the West Midlands. These wards were the same as those used to develop the ICECAP-A measure and were chosen for their socio-economic diversity. The individuals that responded to the invite were then added to those individuals who had earlier expressed an interest in taking part. The resulting group was then purposively sampled on the basis of responses to the screening questionnaire to

[1] 200 each from wards in the 3rd and 4th quartile to allow for lower response rates.

ensure diversity in terms of age, sex, ethnicity, health and socio-economic status. Interviews were conducted until no new substantive issues were emerging in the interviews (i.e. sample size was determined qualitatively).

2.3 Interview conduct

All but one participant was interviewed in their own home. The interview began with a recap of the study aims and an explanation of the format of the interview. The process of thinking-aloud was likely to be unfamiliar to most people and therefore all participants then performed two 'warm up' think-aloud tasks. In the first task they were asked to count up the number of windows in their home (5), thinking out loud as they went. In the second task participants were asked to complete either a single five-point question on their general health *or* a single five-point question on their satisfaction with life.

Each participant then received both the ICECAP-A capability measure and the EQ-5D health measure and was asked to complete them thinking out loud as they went. The order they received the two questionnaires depended on whether they had completed the life satisfaction or health warm-up question first. To prompt them to think-aloud, a standard protocol, based on one used by Gilhooly and Green (1996) (20), was then read out, requiring participants to verbalise their thoughts during the completion of the measures. Participants were not interrupted during the completion of the measure and if they were silent for a period of 10 seconds or more (which was very rare) they were asked to keep thinking aloud.

During the think-aloud task, digital recording was supplemented by written notes on any problems encountered and issues raised, while completing the measures. Following the think-aloud interviews a semi-structured follow-up interview was conducted. The interview started with discussion, using the written notes, to clarify participants' feedback whilst completing the measures. After this, a set of standard questions (e.g. Did you find any questions off-putting? How easy or difficult did you find answering the questionnaires overall?) were used to discuss the participant's experiences and opinions more generally with regard to the questionnaires.

2.4 Interview analysis

All interviews were transcribed verbatim. Analysis of the think-aloud portion of the interviews has been completed and some qualitative insights drawn. Qualitative analysis of the semi-structured follow-up portion is yet to be completed and we would welcome any comments or advice on issues to focus on in this aspect of the analysis.

The think-aloud portion of each interview was divided into 11 segments; 5 representing the items on the ICECAP-A measure and 6 representing the items on the EQ-5D (including the visual analogue scale question). Four raters (the listed authors) then coded the transcripts with the aim of identifying segments of the interview where the participant encountered a problem in the process of completing the question. It has been proposed that the process of completing a survey question can be broken down into four stages: comprehension, retrieval, judgement and

response (16). Participants are required to: (i) understand (comprehend) the question in the way that the researcher intended; (ii) successfully retrieve the appropriate information to answer the question from their long-term memory; (iii) correctly judge how the recalled information should be used to answer the question and (iv) format the information into a valid response for the questionnaire. To try to standardise the coding of transcripts, examples of potential segments that would and would not constitute errors at each stage were drawn up. Each rater then coded the segments in each transcript as either: (a) error-free, (b) containing one or more errors or (c) as a 'struggle'. The latter category was used to identify segments where the participant clearly had difficulty answering the question, but provided a valid answer. Consistency between raters on the coding of the data was then calculated using raw agreement and kappa statistics (21).

Following the independent coding, segments were categorised as errors (struggles) if a majority of coders noted a specific type of error (or struggle). Segments where a majority of coders noted no error of any type or struggle were classified as error-free. Segments where coders were evenly divided, or more commonly a majority noted an error but with no majority on the type of error, were discussed and a code agreed upon by the research team.

A brief description of the completion problems, illustrated by selected quotes, is provided. Unlike most qualitative papers, quotes are not intended to be illustrative of typical responses, instead they are highly selected to illustrate the few occasions that participants made errors or struggled in the think-aloud. In some cases the follow-up semi-structured interview data has been used to expand on the think-aloud response. This will be extended for the full paper.

3. Results

The invites generated 51 responses (75 when supplemented with the responses from the previous study). From this sample frame, 34 interviews were conducted in November and December 2010, lasting between 15 and 60 minutes. The full characteristics of the interview sample are shown in appendix B. In summary, 5 of the 34 participants were non-native English speakers, 11 were over 65 (8 were under 45), 7 were from the most socially deprived ward and the participants had a mean EQ-5D VAS score of 77.6 (std dev 17.8).

3.1 Inter-rater agreement on error codes

Inter-rater agreement was assessed by determining the degree to which the raters agreed on whether a segment was: a) an error or struggle, or b) error or struggle-free. Table 1 shows the level of agreement and kappa coefficients (chance-corrected agreement) between raters for the ICECAP-A and EQ-5D.

Table 1: Agreement between raters on the presence of error or struggle

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1		92% (0.47)	84% (0.40)	92% (0.33)
Rater 2	91% (0.58)		83% (0.36)	89% (0.21)
Rater 3	80% (0.29)	79% (0.31)		81% (0.22)
Rater 4	87% (0.40)	88% (0.53)	81% (0.38)	

Notes:
1. kappa coefficients are in parentheses
2. ICECAP-A agreement data highlighted in bottom left, with EQ-5D agreement data unhighlighted

The level of agreement on the codes applied was similar on both the ICECAP-A (79-91%) and EQ-5D (81-92%). As most segments were coded as without error, a high level of agreement would be expected. A more informative metric therefore is the chance-corrected level of agreement (kappa coefficient) (21). On ICECAP-A, the kappa coefficients for agreement vary between 0.29 and 0.58. On the EQ-5D they vary between 0.21 and 0.47. Using Landis and Koch's (1977) guidelines (22), these levels of agreement would be termed 'fair' (0.21-0.4) to 'moderate' (0.41 to 0.6).

3.2 Completion of the ICECAP-A measure

Table 4 shows the agreed errors and struggles on ICECAP-A, following discussion and agreement on the codes to apply to segments where there was no clear majority. Six (3.5%) out of the 170 segments of the ICECAP-A were associated with an error and 10 (5.9%) with a struggle. Of the errors, three were response errors, two were comprehension errors and one, a judgement error.

Stability (ability to feel settled and secure)

Participants encountered few problems with providing an answer to the stability question, thinking, as intended, about the degree to which a range of factors, such as finances, family and their attitudes, might restrict their ability to feel settled and secure in life. The single response error resulted from one participant providing a tick between two boxes on the measure. The single struggle occurred when a participant found the question vague, but opted in the end for the answer he considered most appropriate:

About feeling settled and secure … a little bit wishy washy there I think. All, many, a few, any. So I think really people would be either settled and secure, settled, secure, is it the same thing, that's the question. You could be settled somewhere but are you secure? … However I will put I am secure in all areas of my life. [TA49]

Table 2: Errors and struggles on the ICECAP-A capability measure

	Stability	Attachment	Autonomy	Achievement	Enjoyment
TA01					
TA02					
TA03		Struggle			
TA08		Struggle			
TA10					
TA11					
TA17					
TA22			Error – Rs		
TA26					
TA31					
TA32					
TA33				Struggle	
TA36				Struggle	
TA37		Error - C			
TA41				Struggle	
TA43			Struggle		
TA45					
TA46			Error – Rs		
TA49	Struggle	Error - J			
TA50					
TA52					
TA54					
TA55					
TA58				Error – C	
TA60					
TA61					
TA62			Struggle		
TA64					
TA65					
TA66					
TA67		Struggle	Struggle		
TA68					
TA69	Error - Rs				
TA73					

Key: Errors: Rs (Response), C (Comprehension), J (Judgement)

Attachment (ability to have love, friendship and support)

The attachment question caused more problems than the stability question, with a number of participants suggesting it was 'ambiguous' or 'odd'. The problems mainly seemed to relate to the capability wording of the question. Participants questioned whether asking them whether they "can have a lot of love friendship and support" referred to whether they could 'deal' with, were 'worthy' of (see below), or had the 'capacity' to have love, friendship and support in their life. Where problems were encountered participants tended to react by reinterpreting the question as asking about functioning:

… I can have a lot of love, friendship and support, I can have quite a lot …I don't really understand what the word can is doing in this question, I don't know whether it

means that I am worthy of having a lot of love, friendship and support or what but assuming that it means that I do have a lot of love, friendship and support, I'd say that I have quite a lot of love, friendship and support. [TA37]

Autonomy (ability to be independent)
In general the autonomy question was quite easily understood, and although there were a couple of problems, these did not appear to relate to the autonomy question in particular (one participant ticked a box that contradicted their verbal response and one wanted to tick between the two middle options). However, one issue was a need to establish the precise meaning of being 'independent', with participants enquiring whether this referred to independence in terms of finances, emotions or physical health, for example:

Again independent meaning...meaning what? Financially, mentally? Physically? If it's mental or physical I'm totally independent, if it's financially I am married so you're never completely independent so I'll go for many things. I'm not sure what it's about. [TA67]

Achievement (ability to achieve and progress)
The achievement question posed a problem for a few older participants because they interpreted the question in terms of their career progression and qualifications. For some participants, near the end of their career or in retirement, the question appeared to have little relevance:

Don't really understand achievement and progress at all, I don't know what you're actually asking for there. At 56, my progress career wise is sort of limited. Achievements, I'm not going to get any more qualifications. I would have to say I pass. Being a go-getter I'm going to say I can achieve and progress in all aspects of my life, but I find that question quite ambiguous. [TA58]

Enjoyment (ability to have enjoyment and pleasure)
The enjoyment question posed no notable problems, with participants having little problem with the concepts. As with other questions, some participants still deliberated and displayed some uncertainty over their response. However, participants all reached an answer that appeared to be consistent with their reasoning and the intended question.

I can have quite a lot of enjoyment? Yes, heavens I can have a lot of enjoyment and pleasure. [I] have reasonable health, reasonable income… yes, I still think I can have a lot of enjoyment and pleasure. [TA03]

Distinguishing between capability and functioning
In the think aloud exercise, most participants read the questions out loud. Some read out the questions as asking about capabilities, as intended, but some also read out the questions as asking about functioning. Furthermore, others read out the questions as a mixture of the two. Participants reading, and appearing to answer, questions in terms of their functionings, for example, "I am independent in many

things" were not categorised as producing an error. In many cases functioning and capability will coincide, either logically where functioning at the top level is indicated (as recognised by TA02 below), or by virtue of a participant functioning at the highest level their capability permits (e.g. TA32, below).

Well I'm able to feel I am, I am, I do feel actually, so I am able to feel settled and secure in all areas of my life basically because the insecurities I have are dealt with by my sister, which is good. [TA02]

I can have a lot of love, friendship and support. I can and do have a lot of friendship and support. [TA32]

One participant had difficulty understanding why someone would indicate that they had anything other than the maximum level of capability for 'enjoyment' or 'attachment', since all people ultimately had the potential for this:

I think this is very similar to number two [attachment]. If the phrasing is how much can you have of enjoyment and pleasure. Can't see why you should have less… I'll go for 4 [top level]. [TA67]

In the interviews, there were also examples where participants struggled to relate the capability concept to their own life. They tended to respond by answering the question as if it was only asking about their functioning, most notably on the 'attachment' question (see earlier). However, even on the attachment question participants were able to distinguish between capability and functioning and suggest a level of capability above what they appeared to be functioning at:

I can have a lot of love, friendship and support… I want to pick and choose when…I don't crave people's company all the time. [TA64]

One participant confirmed, in the post think-aloud discussion, that she noted the capability wording and would probably have answered the question differently had it been worded in terms of functioning:

TA67: … the wording means that if you said I have a lot of love, people would answer something totally different but the can makes it yes please.

HA: Would you have answered differently if it had said I have a lot of love, friendship and support, I have quite a lot?

TA67: Yes, probably. Probably.

HA: How would you have answered?

TA67: I would probably have chosen the second one. Because you get quite a lot of support, friendship, etc, but it's not like as a 10/10. And I think the same for number four. 'Can' makes it potential so you score higher.

As noted earlier, the capability wording appeared most problematic on the attachment attribute, however on other attributes, notably autonomy and achievement it appeared to be more easily interpreted, with a number of participants distinguishing their capability from their functioning:

I would be able to be independent in many things but I am probably only independent in a few things. That is most probably my choice but, as I said, I have a two year old child and I run my own business… [TA46]

So in a way I think rethinking about it I can achieve progress in a few areas. Well no I suppose, if I was really intent and focused I could achieve some of the others…Yes I think I can, its my apathy that probably, its not an impossibility to improve progress in many aspects of my life. Its, so yes in, upon reflection I'll go for I can achieve many aspects. [TA03]

3.3 Completion of the EQ-5D health measure

Table 3 shows the final error and struggle pattern on the EQ-5D, following discussion and agreement on the codes to apply to segments where there was no clear majority. Twelve (5.9%) out of the 204 segments of the EQ-5D were associated with an error and three (1.5%) with struggle. Of the errors, 11 were response errors and one was a comprehension error.

Mobility

Although most participants found the mobility question simple and easy to complete, three participants were coded as making response errors. In two cases this was because the participant had suggested verbally that they had some problems walking about, but ticked the 'no problems' box on the measure.

When pressed on the issue in the follow-up interview afterwards, one participant revealed that their mobility problems occurred at the weekend and so overall they felt like 'no problems' was the correct answer:

…. I think overall, I had quite a bad accident in my mid teens, I broke both my hips and both my legs and in the winter I do get a bit of arthritis and things because of it …I only put the problems down to … how I stand at a till all day at the weekend, 'cos Monday to Friday I don't have any problems so I think standing up all day just doesn't help that … so I just think overall that [top box] would be correct … [TA22]

Self care

Just a single problem was encountered on the self-care question. Here, one participant ticked between the middle and upper levels. This participant, who at 91 was the oldest participant in the sample, also made a similar error on the stability question on ICECAP-A.

Table 3: Errors and struggles on EQ-5D

	Mobility	Self-care	Usual activities	Pain	Anxiety/ Depress	VAS
TA01	Error - Rs					
TA02				Error - Rs		
TA03				Error - Rs	Error - Rs	
TA08						
TA10				Error - Rs		
TA11				Struggle		
TA17						
TA22	Error - Rs					
TA26						
TA31						
TA32						
TA33						
TA36						Error - Rs
TA37						
TA41						
TA43			Error - C		Struggle	Error - Rs
TA45	Error - Rs					
TA46						Error - Rs
TA49						
TA50					Struggle	
TA52						
TA54						
TA55						
TA58						
TA60						
TA61						
TA62						
TA64						
TA65						
TA66						
TA67						
TA68						
TA69						
TA73		Error - Rs				

Key: Errors: Rs (Response), C (Comprehension)

Usual activities

Similarly, just a single problem was encountered on the usual activities question. In this case, the participant was confused by the term usual activities, unsure whether the reference point should be what was usual for him or what was might be considered 'average' in a more general sense:

I find it a bit confusing saying, 'Usual activities' because it's usual in what sense? In the sense of average, or usual to the limits that I have had to put on my activities? So I would say with my usual activities I have no problems. [TA43]

Pain

There were two distinct types of problem encountered in this sample with the pain question. The first was that some participants had a level of pain that they would place in between 'no pain' and 'moderate' pain, for example one participant talked about their 'occasional moderate pain':

Pain and discomfort. Well there's a big difference between pain and discomfort. There's no scale here so if it was on a scale of say nought to two or three to five, or whatever, the extreme pain would have to be obviously at the upper end. So it's actually quite a difficult thing to say because sometimes it depends, my knees are a bit tricky in the winter, but I wouldn't say I have extreme pain or discomfort. But if it said occasional or on a scale of or how often, I don't know. So I don't have any pain or discomfort from that point of view then because it's not chronic. If it said chronic pain then no. If it says moderate pain occasionally then maybe, though I have no pain or discomfort is the one that would fit in best with that. [TA02]

The second issue appeared to be one of adaptation, where participants recognised that they may be experiencing what others would consider to be pain, but because they could tolerate it, they were reluctant to classify themselves as having pain:

I have no pain and discomfort. I have moderate pain and discomfort, I am in extreme pain. I don't know about that. I don't know what you'd describe as pain because I've been told I have a very, very high pain threshold so I have no pain or discomfort. [TA11]

This comment was explored further in the follow-up interview and the participant revealed that he was less likely to complain about aches and pains than some of his friends:

Well when you speak to people and you mix with friends or relations people complain about various things. I don't feel in any way that I have anything that I have to complain about as far as my pain threshold. If I cut myself or scratch myself or knock myself or whatever, it goes away. So it doesn't worry me unduly. …We have a friend who's down south …she's very cheerful in her attitude but she complains … that a cold is life threatening. [TA11]

Anxiety and depression

Although most participants appeared to be able to answer the anxiety and depression question, a couple of participants struggled with the wording of the levels. One participant wondered just how anxious or depressed they had to be in order to be classed as 'extremely' anxious. Similarly, another participant, a non-native English speaker, struggled with the term 'moderately anxious or depressed', wanting clarification:

I'm moderately anxious and depressed. What is the definition of moderately anxious and depressed? [Interviewer: I can't really interrupt you] Alright, no but I'm asking

you a question, I'm going to complete it myself. But I am asking you to help my defining the answer… [TA50]

Visual analogue scale
Two response errors occurred on the VAS because participants drew lines that intersected the rating scale in multiple places. In one case this was because a participant appeared to correct their initial response, but with the result that a line was drawn that crossed the rating scale at both 90 and 95. In the other case, the participant was not able to combine his rating of his mental health with that of his general health and so chose to represent both on the VAS:

Okay, well I'm thinking of breaking it up again as I did in the first [practice] question, and I'm going to write underneath 'general' and 'mental', from there I'll do the arrows as stated to the points that I think suit … [TA43]

4. Discussion

To ensure meaningful data for economic evaluation, it is important to know whether self-reported quality of life questions are being interpreted in the way that they are intended to be. This study found that individuals generally interpreted questions in the intended manner and encountered problems (of any sort) on fewer than 10% of questions across the EQ-5D and ICECAP-A measures. For the EQ-5D, an important issue is the degree to which individuals' desired response category is available. For the ICECAP-A, individuals sometimes struggled to clarify the meaning of a broad concept (such as being 'independent') and the meaning of the capability concept.

The pattern of response errors on the EQ-5D suggests that individuals were generally able to understand, retrieve information and make a valid judgement in response to the questions, but then sometimes ran into difficulty formatting the information for the questionnaire. In particular, some participants perceived a big gap between no pain and moderate pain and this caused a problem if they had some pain but did not consider it to be 'moderate'. In general, participants tended to round down and were more likely to indicate 'no problems' rather than 'moderate problems' when faced with low levels of, or transient, pain. Specifically this study suggests that the current EQ-5D is likely to be categorising a number of patients with more minor health problems into the top states. From the table in appendix B, it can be seen that 10 of the 11 individuals that classed themselves into the top health state reported VAS scores below 1 (mean 0.86). This issue is likely to be addressed to some degree by the new 5-level version of the EQ-5D (23). The new measure, for example, allows individuals to express 'slight' pain or discomfort, and though not explored in these interviews, it seems likely that a number of individuals who choose 'no pain' would have chosen 'slight pain' had the option been available.

For the ICECAP-A, the pattern of response errors and struggles suggests that more problems were encountered in the initial comprehension stage. Although complete misunderstandings of terms were rare, general terms such as 'independent', 'secure'

or 'achieve' often prompted participants to speculate on what these meant in the context of the questionnaire. The *literal* meaning seemed to be there, but there was sometimes doubt about the *intended* meaning. One option to address the perceived subjectivity would be to provide more detailed information about the meaning of the terms within the questionnaire. However, caution is required as additional terms potentially invalidate value sets calculated using the original, more parsimonious, descriptive system. Furthermore, individuals can latch onto certain terms in descriptive systems (10;11) and illustrative terms can lead to confusion if they are not seen as relevant to individuals' lives (4). As a result, extra descriptive information may actually distract respondents from thinking about the broader meaning of the question and overly focus on specific aspects of the question. In this study, where participants articulated what they were thinking, general lay understandings closely matched the concepts embodied in the broad terms used. For example, participants identified that independence could cover financial independence, emotional independence and physical independence, amongst other things. This suggests that even if individuals take slightly longer to respond to wellbeing questionnaires, containing broader constructs than health questionnaires, their answers are no less valid.

One ICECAP-A question that caused problems was 'achievement'. Younger participants and some older participants were fine with the question and clearly identified ways in which they could, or could not, achieve in life. However some older participants interpreted the question solely in terms of career and qualifications and, for them, this rendered it slightly irrelevant. For other older participants, there was a degree of reinterpretation needed with some starting to interpret achievement less in terms of career success and more in terms of their health and hobbies. This may be a form of response shift (7) with the participants reconceptualising the meaning of a term over time. A more detailed analysis of these segments is needed to understand how these responses may be related to adaptation issues.

Although it has been argued that capabilities are, philosophically, a more desirable interpretation of wellbeing than functionings (24), most work to operationalise the capability approach has focused on measuring functionings (25). A key objective of this study, therefore, was to explore the degree to which individuals can respond to self-complete questions about their capabilities. The think-aloud findings show that many participants clearly grasped that the questionnaire was asking them to indicate the degree to which they were *able* to function. In a number of cases they emphasised this by revealing that although they were capable of attaining a higher level of functioning, for one reason or another they functioned at a lower level. This occurred, for example, because of apathy (in the case of achievement), family commitments (in the case of independence) and wanting to 'pick and choose' (in the case of attachment). Clear response in terms of capability was by no means observed across the sample: many participants read the questions in terms of their functioning and/or articulated clear difficulties with the capability concept, most notability on attachment. It may be that individuals struggle with the term "I can" as indicating capability. Although the dictionary definition of "can" is synonymous with the idea of

capability[2] and "can" appears to be an appropriate term (10;11), the word did seem to result in more divergent interpretations in this think-aloud exercise. This raises the possibility that a single way of communicating the capability concept in the measure, along with perhaps, explicit instructions at the start, may help reinforce that fact that the questions are about capabilities, rather than functionings.

An interesting issue is whether responses to the ICECAP-A or EQ-5D seemed to be more subject to adaptation. Other than concept reinterpretation of 'achievement' (discussed previously) there was little evidence of adaptation on the ICECAP-A from the think-aloud protocol. Although some older participants struggled with the terms 'achieve and progress' there was no evidence of a lowering of standards over time. Conversely, there did seem to be clearer evidence of adaptation on the EQ-5D, with participants using their own lower/higher internal standard for benchmarking their level of 'usual activities' or 'pain'. This is clearly an area for further qualitative investigation within this study.

The use of think-aloud exercises to explore the completion process for surveys is not without potential pitfalls. Willis (2005) suggests participants may stray from the task, fail to provide think-aloud streams (i.e. directly retrieve information) and produce responses that are biased by the process of thinking aloud (5). Whilst, there were a few examples of individuals 'straying' or engaging in 'direct retrieval', in general, the experience of using think-aloud in this study was quite positive. Whilst many participants were hesitant at first, the warm up window counting exercise encouraged participants to feel that they could think-aloud. Participants then tended to engage in a pattern of reading out the question, and then explaining the rationale for their response as they went, occasionally pointing out things that struck them as odd about the question. In a few cases, participants indicated that their response process under different conditions may have been slightly different: one individual for example suggested he would answer with greater suspicion if being asked the questionnaire in a clinical setting and another suggested she might take more time over the questionnaire had it arrived through the letterbox. In general, however, the think-aloud process yielded useful insights in terms of the understanding of the questions and the source(s) of problems.

5. Conclusion

Both questionnaires appear to be easily understood by the majority of participants, although specific problems were encountered occasionally on both questionnaires. The five-level version of the EQ-5D is likely to reduce the numbers of individuals who

[2] 1a. Used to indicate physical or mental ability
1b. Used to indicate possession of a specified power, right, or privilege
1c. Used to indicate possession of a specified capability or skill
2a. Used to indicate possibility or probability
b. Used to indicate that which is permitted, as by conscience or feelings
c. Used to indicate probability or possibility under the specified circumstances
Source: http://www.thefreedictionary.com/can

want to tick between the boxes on the EQ-5D, but is not likely to reduce the numbers of individuals who use an internal reference standard when answering the questions. On ICECAP-A, although providing more detail may remove some of the perceived subjectivity of the questions, this would need to be done in a way that did not invalidate the value sets or lead to confusion. Individuals also appear to be able to identify where their capability and functioning differs and report capability but this was by no means the case for all individuals. The degree to which individuals can report capability information and the reasons why functioning and capability may diverge still requires further examination. In summary, there are some issues with the both questionnaires and these should be considered when developing/refining self-report capability, wellbeing and health measures. Nevertheless the relative lack of difficulty that individuals had in completing both questionnaires is encouraging in terms of their use with the general adult population in the UK.

Some points we would welcome views and discussion from HESG on are:

- The issues that should be analysed in more detail in the qualitative work
- The degree to which the completion problems encountered are concerning for outcome measurement
- Any views on particular problems, and advantages, of using the measures (especially ICECAP-A) in certain population/patient groups

Appendix A: the ICECAP-A questionnaire

ABOUT YOUR OVERALL QUALITY OF LIFE

Please indicate which statements best describe your overall quality of life at the moment by placing a tick (✓) in ONE box for each of the five groups below.

1. Feeling settled and secure		
I am able to feel settled and secure in all areas of my life		4
I am able to feel settled and secure in many areas of my life		3
I am able to feel settled and secure in a few areas of my life		2
I am unable to feel settled and secure in any areas of my life		1

2. Love, friendship and support		
I can have a lot of love, friendship and support		4
I can have quite a lot of love, friendship and support		3
I can have a little love, friendship and support		2
I cannot have any love, friendship and support		1

3. Being independent		
I am able to be completely independent		4
I am able to be independent in many things		3
I am able to be independent in a few things		2
I am unable to be at all independent		1

4. Achievement and progress		
I can achieve and progress in all aspects of my life		4
I can achieve and progress in many aspects of my life		3
I can achieve and progress in a few aspects of my life		2
I cannot achieve and progress in any aspects of my life		1

5. Enjoyment and pleasure		
I can have a lot of enjoyment and pleasure		4
I can have quite a lot of enjoyment and pleasure		3
I can have a little enjoyment and pleasure		2
I cannot have any enjoyment and pleasure		1

Please ensure you have only ticked ONE box for each of the five groups.

Appendix B: Characteristics of the interview sample

ID	Sex	Age group	English 1st lang	Electoral ward	ICECAP-A responses					EQ-5D responses					
					Sta	Att	Aut	Ach	Enj	Mob	S/C	U/A	Pain	Anx	VAS
TA01	M	65+	Yes	1st quartile	3	2	4	3	4	1	1	1	2	2	90
TA02	F	45-64	Yes	2nd quartile	4	4	4	3	4	1	1	1	1	1	92
TA03	M	65+	Yes	3rd quartile	3	3	3	3	4	2	1	1	2	1	91
TA08	F	45-64	Yes	3rd quartile	3	4	4	3	4	2	1	2	2	1	80
TA10	F	45-64	Yes	3rd quartile	3	4	3	3	4	1	1	2	1	1	60
TA11	M	65+	Yes	2nd quartile	3	4	4	3	3	2	1	2	1	1	85
TA17	F	45-64	Yes	4th quartile	4	4	4	4	4	1	1	1	1	1	92
TA22	F	<45	Yes	1st quartile	2	2	2	1	2	1	1	1	2	2	50
TA26	F	<45	Yes	1st quartile	3	4	3	3	4	1	1	1	1	1	95
TA31	M	<45	Yes	3rd quartile	3	4	3	3	4	1	1	1	1	1	92
TA32	F	45-64	Yes	3rd quartile	4	4	4	3	4	1	1	1	1	1	95
TA33	M	45-64	Yes	3rd quartile	3	4	2	3	4	2	1	2	2	2	30
TA36	M	65+	Yes	1st quartile	3	4	4	3	4	1	1	1	2	1	90/95
TA37	F	45-64	Yes	1st quartile	3	3	4	3	3	1	1	1	2	2	60
TA41	M	65+	Yes	2nd quartile	3	4	2	2	2	2	1	2	2	1	70
TA43	M	45-64	Yes	3rd quartile	1	2	3	1	2	1	1	1	2	3	70/17
TA45	F	65+	Yes	3rd quartile	4	4	4	3	4	1/2	1	1	2	1	90
TA46	F	<45	Yes	3rd quartile	3	4	3	3	4	1	1	1	1	1	
TA49	M	45-64	Yes	2nd quartile	4	4	3	4	4	1	1	1	1	1	90
TA50	M	65+	No	4th quartile	2	2	3	3	3	2	2	2	2	2	
TA52	M	45-64	Yes	4th quartile	4	4	3	4	4	1	1	1	2	1	99
TA54	M	45-64	Yes	1st quartile	3	4	3	3	4	1	1	1	1	1	90
TA55	M	65+	Yes	4th quartile	3	4	3	3	4	1	1	1	2	2	90
TA58	F	45-64	Yes	1st quartile	4	4	4	4	4	1	1	1	2	1	75
TA60	F	65+	Yes	2nd quartile	4	3	4	3	3	1	1	1	2	1	77
TA61	F	<45	Yes	1st quartile	4	4	3	3	3	1	1	1	1	1	100
TA62	M	<45	Yes	1st quartile	3	4	2	3	3	1	1	1	1	1	60
TA64	F	65+	No	4th quartile	3	4	3	3	3	1	1	1	2	2	75
TA65	M	<45	No	4th quartile	2	4	2	3	2	1	1	1	2	2	60
TA66	M	45-64	No	4th quartile	3	3	2	3	2	2	2	2	2	2	70
TA67	F	45-64	No	2nd quartile	3	4	3	3	4	1	1	1	1	1	70
TA68	F	45-64	Yes	2nd quartile	3	4	4	3	4	1	1	1	1	1	75
TA69	F	<45	Yes	1st quartile	3	4	4	3	4	1	1	1	2	1	85
TA73	F	65+	Yes	2nd quartile	4/3	4	3	1	3	2	1/2	2	2	2	40

Reference List

(1) Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.

(2) Brazier J, Jones N, Kind P. Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. Quality of Life Research 1993;2:169-80.

(3) Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics 2004;13:873-84.

(4) Mallinson S. Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. Social Science and Medicine 2002;54:11-21.

(5) Willis G. Cognitive interviewing: a tool for improving questionnaire design. Sage Publications: Thousand Oaks, CA; 2005.

(6) Menzel P, Dolan P, Richardson J, Olsen J-A. The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. Social Science and Medicine 2002;55:2149-58.

(7) Sprangers M, Schwartz M. Integrating response shift into health-related quality of life research: a theoretical model. Social Science and Medicine 1999;48:1507-15.

(8) Coast J, Smith R, Lorgelly P. Should the capability approach be applied in health economics. Health Economics 2008;17:667-70.

(9) Dolan P, Kahneman D. Interpretations of utility and their implications for the valuation of health. The Economic Journal 2008;118:215-34.

(10) Al-Janabi H, Flynn T, Coast J. Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. Quality of Life Research 2011;In press.

(11) Coast J, Flynn T, Natarajan L, Sproston K, Lewis J, Louviere J, et al. Valuing the ICECAP capability index for older people. Social Science and Medicine 2008;67(5):874-82.

(12) Grewal I, Lewis J, Flynn T, Brown J, Bond J, Coast J. Developing attributes for a generic quality of life measure for older people: preferences or capabilities? Social Science and Medicine 2006;62:1891-901.

(13) Brooks R. EuroQol: the current state of play. Health Policy 1996;37:53-72.

(14) Ryan M, Watson V, Entwistle V. Rationalising the 'irrational': a think aloud study of discrete choice experiment responses. Health Economics 2009;18(3):321-36.

(15) Westerman M, Hak T, Sprangers M, Groen H, van der Wal G, The A. Listen to their answers! Response behaviour in the measurement of physical and role functioning. Quality of Life Research 2008;17:549-58.

(16) Collins D. Pretesting survey instruments: an overview of cognitive methods. Quality of Life Research 2003;12:229-38.

(17) Ericsson A, Simon H. Verbal reports as data. Psychological Review 1980;87(3):215-51.

(18) Kuusela H, Paul P. A comparison of concurrent and retrospective verbal protocol analysis. The American Journal of Psychology 2000;113(3):387-404.

(19) Brod M, Tesler L, Christensen T. Qualitative research and content validity: developing best practices based on science and experience. Quality of Life Research 2009;18:1263-78.

(20) Gilhooly K, Green C. Protocol analysis: practical implementation. In: Richardson J, editor. Handook of qualitative research methods.Leicester: BPS books; 1996. p. 55-74.

(21) Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

(22) Landis R, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-74.

(23) Janssen M, Birnie E, Haagsma J, Bonsel G. Comparing the standard EQ-5D three-level system with a five-level version. Value in Health 2008;11(2):275-84.

(24) Sen A. The idea of justice. London: Allen Lane; 2009.

(25) Chiappero-Martinetti E, Roche J-M. Operationalization of the Capability Approach, from theory to practice: a review of techniques and empirical applications. In: Chiappero-Martinetti E, editor. Debating global society: reach and limits of the capability approach.Milan: Fondazione Giangiacomo Feltrinelli; 2009. p. 157-201.