Extending the QALY: generating, selecting and testing items for a new generic measure of quality of life – preliminary results

John Brazier, on behalf of the E-QALY team and international collaborators

Abstract

BACKGROUND: The Extending the QALY (E-QALY) project aims to develop a broad generic measure of quality of life for use in economic evaluation across public sectors with a key focus on health, social care and public health, based on the views of users and beneficiaries of these services. A targeted qualitative review identified seven high level domains (with sub-domains): feelings and emotions, cognition, activity, self-identity, relationships and social connections, 'coping, autonomy and control' and physical sensations. This paper will present the preliminary results of the next two stages of the project: (1) generating and selecting a pool of items for testing, and (2) undertaking cognitive interviews to establish the face validity of those items.

OBJECTIVES: The aim was to generate, select and qualitatively test the items for the new measure.

METHODS: The first step in item generation was to draw structured information from the qualitative literature review. The terminology and concepts associated with each sub-domain was identified from the review, with a particular focus on the language used by respondents. After group discussions the terms/concepts within each sub-domain were reviewed by one team member, and checked/challenged by one other. Where appropriate, this led directly to new items for consideration.

Secondly, items from 30 existing health and wellbeing measures (458 items) and item banks (229 items) that covered the sub-domains were reviewed for potential inclusion based on a set of item selection criteria around: ease of completion (e.g. avoid ambiguity, double barrelled items, double negative wording); ensuring items are not value laden; ensuring good coverage of sub-domains and severity range; ensuring items cover current QoL; suitability for translation; and suitability for valuation. Item wording was refined to meet these criteria and ensure consistency. This process was undertaken using group discussion within the Sheffield research team. Potential recall periods were reviewed and a 7-day recall period adopted.

Semi-structured cognitive qualitative interviews are being conducted in six countries (Argentina, Australia, China, England, Germany and the USA) with individuals with various physical and mental health conditions, carers and social care users to test the content and face validity of the draft items using a standardised protocol.

RESULTS: Items were generated from the two sources identified. Items were then selected through applying the criteria. Items were initially tested with stakeholder consultation i.e. advisory group (n=120), steering group (n=12), PPI group (n=7) and members of the NICE Citizens Council (n=5). An initial set of over 100 items were selected for face validation testing. The cognitive interviews are on-going and we focus on some early results from England.

CONCLUSIONS: Findings from the UK face validity interviews will be assessed in conjunction with findings from face validation in five other countries and additional evidence generated through analysis of the Stage 4 psychometric survey in order to identify preferred items for a future classification system.

Background

Economic evaluation has been adopted around the world to inform the allocation of scarce health care resources. One of the most commonly used methods has been to estimate the incremental cost per Quality Adjusted Life Year (QALY) of new health technologies. QALYs provide a way to capture benefits in terms of impact on survival and health related quality of life (HRQL). HRQoL is valued on a utility scale where one is full health and zero is equivalent to being dead. The QALY principle calculates QALYs gained from an intervention as the value of HRQoL times duration.

EQ-5D is one of the most widely used measure for generating utility values (Wisløff et al, 2014; Richardson et al, 2014). EQ-5D focuses on core health domains (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) which is too narrow even for some areas of health care: indeed, it has been shown to lack sensitivity in a number of clinical areas such as dementia, mental health, hearing and vision (Finch et al, 2017; Brazier et al, 2014; Longworth et al, 2014). The EQ-5D has also been perceived as too narrow to capture benefits in other sectors including social care (defined as "a range of long-term care activities… provided in response to needs arising from physical or sensory impairments, learning difficulties and mental health problems, including those associated with older age", Netten et al 2012:4) and public health. For social care and in some cases health care of long term conditions, the outcomes of care are not only improved health *per se*, but improved quality of life for the recipients from a better meeting of their wants and needs (e.g. nutrition, accommodation, relationships, independence). There are also important consequences for the quality of life of informal (unpaid) carers who support the health care system and provide much of the social care. As a result, other measures have been developed for use in the evaluation of social care interventions, including the Adult Social Care Outcomes Toolkit (ASCOT) (Netten et al, 2012) and the capability measure ICECAP (Coast et al, 2008). For public health there has been an interest in broader wellbeing measures such as the Warwick Edinburgh Mental Wellbeing Scale (WEMWBS) (Tennant, 2007). For carer quality of life, there are a number of measures designed for use in economic evaluation including the CarerQoL-7D (Brouwer et al, 2006) and the Carer Experience Scale (Al-Janabi et al, 2011).

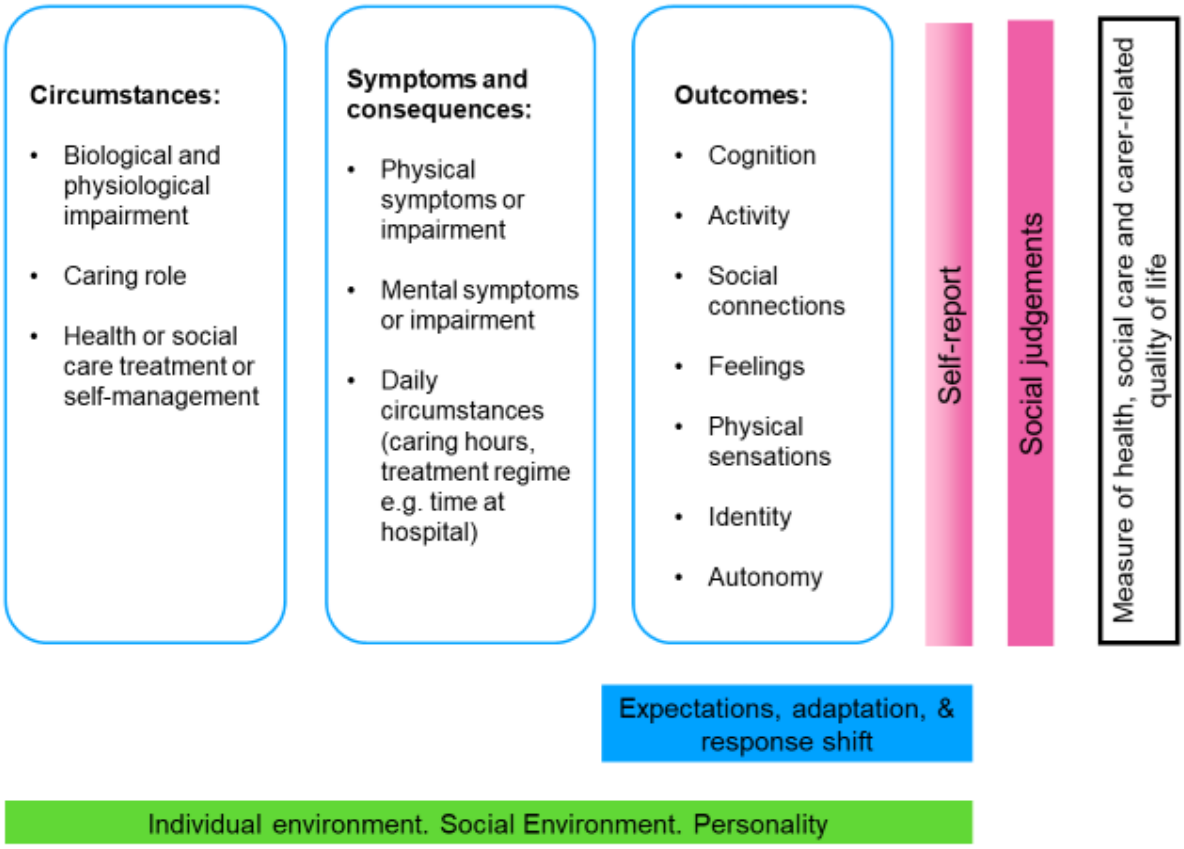
Previously members of the research team had undertaken qualitative work in the UK with decision makers (e.g. NICE Appraisal committee members, NHS-England, Public Health England, local commissioners etc.) and members of the public on their views on the use of wellbeing in the context of health and social care. This work indicated that there are mixed views regarding the use of wellbeing (Peasgood et al., 2016). Although the role for wellbeing-related outcomes was perceived as strongest for social care, carers, mental health and palliative care, across all sectors it was considered at least one of the relevant considerations. Key messages from the participants were that: outcomes such as relationships, a sense of control, being able to do the things you want to, and positive emotion were considered important aspects of quality of life that were relevant to resource allocation decisions; current measures, such as EQ-5D, have insufficient content capturing these aspects; that health (including physical functioning) continues to be important; and decision makers lack the tools necessary to consistently incorporate wellbeing into decision making (i.e. valid, well understood measures). The study identified a strong perceived need for an instrument that captures both health in its broadest sense (physical, mental, emotional and social health) alongside broader key aspects of quality of life (relationships, how we spend our time, control).

The Extending the QALY (E-QALY) project therefore aims to develop a broad generic measure of quality of life that covers both health and broader outcomes and is valued on the zero to one scale necessary to calculate QALYs for use in economic evaluation across public sectors with key focus on health care, social care and public health. The project has six stages with input from a public involvement group (n=7 including patients, carers and members of the general public), a virtual advisory group (n=120 including clinicians,

policy makers, academics and international PROMS experts) and a steering group (n=12).

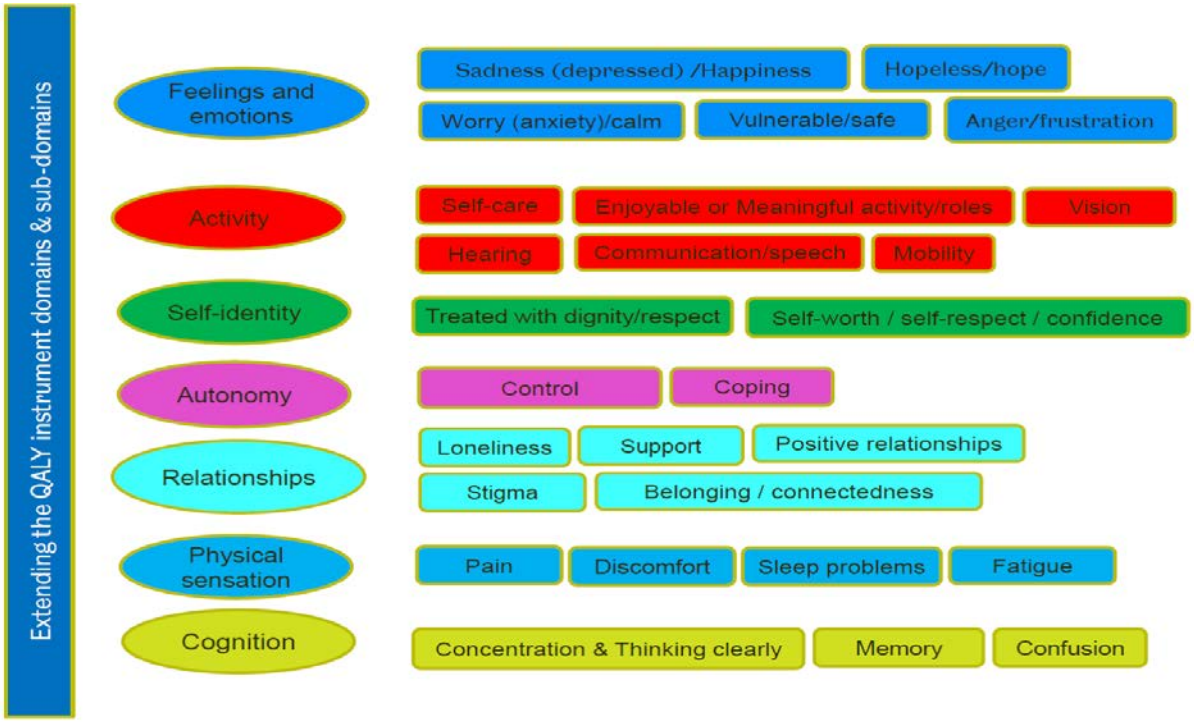
Stage 1 established the domains for the quality of life instrument drawing on a number of strands of qualitative and quantitative work (and reported at the Academy Meeting in March 2018). This was based on a conceptual model (Figure 1) which was used to formulate the extraction framework used in the literature review and to inform the synthesis and subsequent stages of the project. The conceptual model was an extension of the original model developed by Wilson and Cleary (1995) for health to the broader context of the E-QALY project. It starts with the conditions, which may be a biological problem, care intervention or caring role, and moves through to the consequences for functioning and ultimately the impact on quality of life. The conceptual model (Figure 1) has an implicit flow from left to right but does not include arrows because directionality is complex for many of the relationships (e.g. wellbeing domains may impact on health, as well as the other way round). The conceptual model allows the presence of a physical symptom (e.g. vision difficulty) or circumstances (e.g. 24/7 carer) to have a direct impact on quality of life and not just an impact via functioning/activity/social relationships/feelings/identity.

Figure 1: Conceptual framework



A targeted systematic search was undertaken to identify primary qualitative work used in measure development and qualitative reviews on the impact of health conditions, being an informal carer and social care use on quality of life. Framework analysis was used to identify domains and sub-domains. Targeted extraction and synthesis resulted in seven high level themes/domains with sub-themes/domains (Figure 2).

Figure 2: Preliminary domains and sub-domains at the end of stage 1



Stage 2 aims at generating a list of potential items to cover the agreed domains from Stage 1. Stage 3 tests the face validity of candidate items using semi-structured interviews. The findings from Stages 1 to 3 will contribute to a proposed instrument which will be tested and valued during stages 4 to 6. This paper presents work undertaken in Stages 2 and 3 with specific focus on the UK.

1 Method

Item generation, selection and testing is an iterative process informed by best practice as well as consultation with project governance groups including a public involvement group, an advisory group and steering group. The different groups were asked to comment on the methods as presented here, and provided comments on early drafts of the items before they were taken forward for face validation.

1.1 Item selection / generation

The primary goal for the final E-QALY questionnaire is that it can be used to estimate QALYs. This requires a classification system that can be valued using preference elicitation techniques. A secondary aim is to develop a longer version that can be used as a non-QALY profile measure of quality of life. The profile allows greater flexibility in terms of questionnaire length and inclusion of items which may conflict with the QALY model or be potentially problematic to include in preference elicitation tasks. The longer profile also allows for greater nuance on the sub-domains that can be measured. However, it is imperative that the classification can be formed from the longer version and that the classification can be used independently. Item generation and selection reflected these dual aims.

1.1.1 Item selection/generation criteria

Item generation and selection was aimed at identifying those items that reflected the underlying construct of the sub-domain of interest and that were likely to be accurately completed and interpreted in the same way by everyone, whether they are service users completing the questionnaire or members of the general public providing preferences. The methods used for identifying items were based on current accepted best practice in developing quality of life measures (Fayers and Machin, 2016; Butt et al, 2012). The items were initially generated from two sources: (1) drawing concepts and terms from the qualitative review and (2) identifying possible items from existing questionnaires and item banks. Potential terms and items were

then reviewed to ensure coverage of the construct of the sub-domain and choosing items that best meet pre-determined selection criteria. Due to the potentially vast number of existing published items on health and quality of life, and the expansive nature of the literature review, application of the selection criteria began at early screening stages of item generation, so that only reasonably suitable items were taken forward to more formal item selection. Identification, generation and selection of items was an iterative process.

The selection criteria that an item was required to meet drew on existing published criteria (Bradburn et al 2004; Streiner and Norman, 2008) adapted to meet the specific needs of creating a generic health, social care and carer related quality of life preference based measure.

The criteria used are summarized in Table 1. The first twelve criteria relate to ease of completion by ensuring items can be easily understood based on the language and content of the question. Items that are value laden (e.g. items about work or volunteering may imply that these are a good thing for all respondents) were excluded as they may lead people to believe there is a right or wrong way of answering the question which would influence how they answer (criterion 13). The next set of criteria (14-17) relate to coverage across domain and sub-domains. Due to the limited number of items possible within a generic preference-based measure, a single item may be required to cover the full spectrum of a domain. These criteria were used to assess the extent to which this was the case. The next criteria (18-23) relate to the need to avoid items that would be unsuitable for an instrument designed to measure the Q part of the QALY. QALY scores relate to specific periods of time and are assumed to be interpersonally and intertemporally comparable. Consequently, it is important that each item is clearly tapping into the specific time period, and does not rely upon comparisons to other people or other time periods. We also considered criteria to ensure the items are suitable for valuation (24-25) and translation (26). As noted above, there is potential to have a non-QALY profile version of the questionnaire therefore items that did not meet the QALY and valuation criteria were retained in the larger item pool.

Table 1: Item selection and generation criteria

	No	Criteria	Explanation
Ease of completion	1	Reading Level Appropriate	The item should be easy to read and understand. Steiner and Norman (2008) recommend a limit for reading age of 12 years of age. However, a reading age of not more than nine may be more appropriate for general UK audience.
	2	Avoid Ambiguity	Avoid questions which have a potentially ambiguous interpretation, are hard to interpret, lack clarity, are too complex, or to vague.
	3	Avoid questions that are very long	Items should be as short as possible but not too short that it loses comprehensibility. Bradburn et al (2004) notes that ill patients and the elderly may be confused by long complicated sentences.
	4	Avoid double – barrelled question	Double-barrelled items are where two or more questions are asked at the same time and the answers for each may be different. This may also be where two different concepts are compounded e.g. anxiety and depression.
	5	Avoid double-negative	Bradburn et al (2004) notes the importance of avoiding double negatives. This includes considering whether a double negative is created by the choice of response options.
	6	Avoid jargon	The vocabulary should not be technical and should be part of everyday vocabulary.
	7	Avoid terms that are colloquial	Excessively colloquial language may be hard to translate and may not be understood by all. e.g. 'down in the dumps'.
	8	Avoid excessively personal questions	Excessively personal or intrusive items may lead to missing values. e.g. suicide ideation, sexual activity
	9	Avoid questions that are not relevant to all	Avoid items that refer to particular circumstances, situation or lifestyle, such as employment or caring role or presence of a health condition.
	10	Avoid questions which might be ethically inappropriate to ask of all groups	Our Public Involvement group emphasised the fact that items which might leave people in a worse frame of mind after completion should only be used where no alternative is available. e.g. asking overly positive (how satisfied are you with your life?) may be insensitive for those in very difficult circumstances or for those close to the end of life
	12	Avoid items that draw on other knowledge	Items that relate to another piece of knowledge, such as what other people think, may be difficult to complete if the responder is not confident in that knowledge. e.g. "other people care about me" "I am a burden to others"
	13	Value-laden words	Judgmental statements may prejudice the respondent and should therefore be avoided. Tone of the question should be neutral.
Good psychometric properties	14	Avoid items that are too extreme	Items that clearly tap into the extreme end of a sub-domain only would need to be supplemented by other items on that sub-domain. e.g. "I have problems feeding myself" may not be sufficient on its own to identify self-care limitations
	15	Avoid items that are too mild	Items that clearly tap into the mild end of a sub-domain only would need to be supplemented by other items on that sub-domain. e.g. I felt nervous
	16	Avoid items that are too specific	Items that tap into very specific symptoms of a sub-category of people may not adequately capture the sub-domain e.g. hearing voices.

	No	Criteria	Explanation
		to a diagnosis/ condition	
	17	Avoid items for which there may be disagreement about monotonicity	There may be items that are potentially ambiguous as to whether more is always better. e.g. self-confidence, being organised, being satisfied with life
Current quality of life and interpersonally comparable	18	Avoid items that are likely to suffer from Differential Item Functioning (DIF)	DIF identifies sub-groups of people who, despite having the same underlying level of at attribute, answer an item differently. e.g. crying questions can be answered differently between men and women even when they have the same level of depression.
	19	Avoid items that make comparisons over time	Items which ask the respondent to make a comparison to another time period or to 'usual' are not suitable as they depend upon what the past or 'usual' is like for the individual. e.g. 'I'm bothered by things that don't usually bother me'
	20	Avoid items that makes comparisons to other people	Items that make comparisons to other people depend upon who the individual choose to use for a comparison, therefore are in conflict with the need for inter-personal comparability. e.g. 'compared to other people my age' 'I felt just as good as other people'
	21	Avoid items that makes comparisons to expectations	Items which make comparisons to a person's expectations or personal norms again are problematic due to lack of inter-personal comparability. Given the self-complete nature of items this may not be avoided entirely, however, items which are less likely to draw on individual expectations will be preferred.
	22	Avoid items that do not lend themselves to short time periods	Items need to clearly tap into the current situation (as restricted by the time period given within the item). Items will not be suitable if they refer (directly or via the respondent's interpretation) to the recent or distant past (beyond the specified time period), or to the future.
	23	Avoid items that focus on a trait	Items which could be interpreted as referring to a personality trait rather than a current feeling/emotion may risk the response drawing on the situation outside of the specified time period e.g. 'I had a bad temper'
Suitable for valuation	24	It is reasonable to expect trade-off against another sub-domain based on improvement in the item	The domains and sub-domains identified are those that have been reported as being things that matter most to patients, social care users and carers. On that basis it is reasonable to think that those doing the valuation would necessarily be willing to trade of improves in the sub-domain against other sub-domains.
	25	The item does not attribute	Where an item attributes a problem to a particular circumstance e.g. "because of X I am unable to do y" "because of my pain I am unable to see my friends", it is problematic to value due to uncertainty as to whether x (pain) or y (seeing friends) is being valued.
	26	Easily translatable	Avoid items with words that do not translate broadly to other languages and cultures.

1.1.2 Drawing concepts and terminology from the literature review

The first step in item generation was to draw structured information from the qualitative literature review based on the domains and sub-domains. The terminology and concepts associated with each sub-domain was identified from the review, with a particular focus the language that respondents in the qualitative studies had used wherever possible. For example, the following terms and concepts were identified in this way for the sub-domain of 'loneliness': "isolation"; "alienation"; "no-one to talk to"; "aloneness" and "loneliness". This terminology was then reviewed to consider what concepts were appropriate with reference to the selection criteria. For example, the term 'alienation' was considered to be a difficult term but the concept was covered by other terms such as 'feeling left out' or 'excluded'. Other examples are provided in Table 2.

Table 2: Extract from literature review/item selection spreadsheet showing terms and concepts not taken forward

Domain	Subdomain	Term/Concept being reviewed	Reason for not including the term
Relationships	Loneliness	Alienation	Difficult term/concept to understand
	Positive relations	Reciprocity	Not relevant to all – need to have someone close to reciprocate
	Belonging/ Connectedness	Social outcast	Term judgmental and sensitive – and covered in concept of 'not belonging'
Activity	Daily living activities	Coping	Already in sub-domain of domain 'control/autonomy/choice'
	Leisure activities	Loss of Lifestyle	Too broad/abstract a concept – covered by not doing things enjoy/want to do
Feelings/ Emotions	Hope/ Hopelessness	Unfulfilled dreams	Age specific – not as appropriate for young as older
	Anxiety/worry	Paranoia	Condition specific (mental health), difficult term to understand

At the end of this process, item-relevant terms and concepts were left for each sub-domain, some of these terms were treated directly as possible items e.g. "I felt lonely".

1.1.3 Identifying possible items from existing measures

A spreadsheet of items from commonly used generic, carer, social care and mental health quality of life measures was developed. The items were categorized into the domains and sub-domains that had been identified in the literature review. Some items fell into more than one domain. Information on the source, relevant sub-domain(s), original item wording, alternative wording, response options and notes on whether there were known problems with the item such as covering more than one concept was documented. Alternative wording was used to modify items where the original item did not fit the proposed structure or criteria for item selection of the new measure. For example, the item from EQ-5D 'Anxiety and depression' was split into two potential items 'I felt anxious' and 'I felt depressed'.

All items (n=458) from the generic measures listed in Table 3 were included. This included existing preference based measures in health and social care as well as non-preference-based measures that

captured well-being. Item banks and other measures were also screened for items (n=229) which were added to the database (Table 3). After the initial round of item generation and selection, a review was undertaken to identify where selected items did not fully represent the domain or link sufficiently closely to the findings of the literature review. For example, the concepts of 'support', 'stigma' and 'cognition' were identified as being inadequately covered at this stage. Targeted instruments, and a recent study reviewing measures for assessing wellbeing, happiness and quality of life (Linton et al, 2016) were used to help identify more items to address these gaps.

Table 3: Generic and other measures used to identify items

Generic Measures	**PROMIS Item Bank**
17D (Quality of Life in Pre-Adolescence) AQoL-8D (Assessment of Quality of Life) ASCOT (Adult Social Care Outcomes Toolkit) ASCOT-Carer (Adult Social Care Outcomes Toolkit-Carer) CASP-19 (Control, Autonomy, Self-realization, Pleasure) CarerQoL (Carer Quality of Life) CES (Carer Experience Scale) CES-D (Center for Epidemiologic Studies Depression Scale) CIT (Comprehensive Inventory of Thriving) EQ-5D-5L (EuroQoL Health Status Measure) GHQ (General Health Questionnaire) PWS (Psychological Well-Being Scale) ReQoL (Recovering Quality of Life) SF-36 v2 (Short Form-36) WEMWBS (The Warwick Edinburgh Mental Well-Being Scale) WHOQoL (World Health Organisation Quality of Life Assessment Scale) Health Utilities Index	ASCQ-Me® v2.0 Neuro-QOL Item Bank NIH Toolbox Item Bank
	Other Measures
	BBC Subjective Well Being Scale BMPN – Balanced Measure of Psychological Needs Scale ICECAP-O LVQoL Low Vision Quality of Life Scale McGill Quality of Life Questionnaire Nottingham Health Profile QUAL-E: Quality of Life at the End of Life ReQoL (100/1597 items taken to face validity interviews) Rosenberg Self-esteem scale Ryff's Scales of Psychological Well-Being The Short Depression-Happiness Scale (SDHS) The Stigma Scale Visual Functioning Questionnaire
Further measures reviewed	
15D AFFECTOMETER Affect Balance Scale AIMS BPNS Basic Psychological Needs Scale BPSS-Biopsychosocial Inventory CHI Chinese Happiness Inventory De Jong Gierveld Loneliness Scales Duke Social Support Index Emotional Well-Being Scale ENRICHD Social Support Instrument FACT-G Functional Assessment of Cancer Therapy-Cognitive Functioning FS Flourishing Scale HAD Hospital Anxiety and Depression Scale Herth Hope Index ICECAP-O IPPA Inventory of Positive Psychological Attitudes Jarel Spiritual Well-Being Scale	Keller Symptom Questionnaire Kidscreen 52 Life Orientation Test – Revised MOS Social Support Survey MSQoL-54 (Multiple Sclerosis Quality of Life Scale) OHQ Oxford Happiness Inventory PGWB (Psychological General Well-Being Scale) Quality of Life Measures for Nursing Home Residents (Kane et al) QWB-SA Quality of Well-Being - Self Administered Scale Self Esteem Scale SHIS Salutogenic Health Indicator Scale (SHIS) Spiritual Well-Being Questionnaire State-Trait Anxiety Inventory UCLA Loneliness Scale

1.1.4 Selecting the recall period

Recall periods adopted for quality of life vary from today, yesterday, last week (or last seven days), the last two weeks, or last month. The recall period may have an impact on applicability which results in missing items. Very short recall periods such as today/yesterday may mean that respondents are not experiencing the issues raised on the particular day (Bradburn et al, 2004). On the other hand, respondents may not remember information accurately over a long recall period and will only report the most salient information rather than 'on average' (Bradburn et al, 2004). The need to generate an instrument that could be used to track progress following acute events (such as stroke or fracture) in which quality of life may change fairly rapidly, also makes longer periods of time problematic.

A default position of seven days was adopted at the outset, with regular consideration as to whether this would be most suitable for each item.

1.1.5 Selecting the response options

A number of aspects were taken into consideration around choice of response options. This included whether or not frequency or intensity best distinguished the level of attainment for a sub-domain and the specific wording used. The number of levels was also considered based on other instruments, evidence from the literature and judgement within the research team; a default position of five levels was adopted. The levels also needed to be meaningful in the preference elicitation task. Further testing of response choices will also be undertaken in the psychometric analysis.

1.1.6 Questionnaire items for face validation

The final selection of items for the new instrument will not take place until after the face validity interviews in the UK and other countries, and analysis of the psychometric survey. However, consideration of the likely nature of the instrument feeds into the initial choice of items for the face validation.

In many cases, questions could be asked in a positive manner (e.g. I felt happy) or a negative manner (e.g. I felt unhappy). Positive items may be preferred by those in good quality of life while negative items may be preferred by those in poorer quality of life as they reflect their lived experiences. Using both positive and negative items may be confusing to individuals especially in the preference elicitation task e.g. if 'all of the time' is good thing for one item, and not a good thing for another item this is likely to be confusing. In the initial selection, both positive and negative items were included with further consideration on this issue to be undertaken using face validity results and the psychometric survey analysis.

The selection of the number of items included in the item pool is also influenced by the requirements of the psychometric survey. In order to test the domain and sub-domain structure and conduct analysis to explore the performance of items it is necessary to include in the survey sufficient items to enable to the sub-domain to be clearly identified. This requires at least four items on any sub-domain to be tested (Netemeyer et al, 2003) although these can be supplemented in the psychometric analysis from other measures that will be included in the survey e.g. EQ-5D, S-WEMWBS and ASCOT.

There was a prior consultation of the proposed items that involved focus groups with the PI group (n=7) and members of NICE Citizens Council (n=5). The proposed items were presented to the group and participants were asked to share their thoughts on each item. Researchers made notes and the NICE Citizen's Council focus groups was recorded. Advisory group members (n=120) were sent the proposed items via an on-line survey. They were asked to highlight which items they considered problematic and to provide comments to support their choices (n=28 members gave feedback).

1.2 Face validity interviews

In psychometric theory, content validity is an important concept as it refers to how well the items on the questionnaire reflect the scope of the what the questionnaire is trying to measure, or its comprehensiveness (McDowell and Newell, 1996). The face validity, or 'respondent validity' is concerned with how appropriate, relevant and understandable the items on a questionnaire are for the individuals who complete them. It is also important to explicitly check the response levels (or response options) to each question to establish whether participants are able to discern the differences between them. Face validity helps ensure that a measure is acceptable to respondents, which should help with response rates.

A typical way of evaluating the face validity of an instrument is to show the proposed questions to the appropriate groups using cognitive interview methods (Jobe et al, 2003) in order to explore the relevance and meaning of each proposed item. The focus in these face validity interviews is on ease of answering, interpretation and meaning of questions and responses rather than on relevance. All of the sub-domains taken forward to this stage have been identified as important to most groups through the Stage 1 literature review.

1.2.1 Sample

In the UK face validity study, the aim was to recruit social care users (n=10), patients: acute and long term, including the frail elderly (n=10), mental health service users (n=10), carers (n=10) and general healthy public (n=5 to 10). Participants were aged 18 and above, had capacity to consent and were able to read the questionnaire items. Parallel studies are on-going in Argentina, Australia, China, England, Germany and the USA. Here we focus on the initial findings from the UK (further data will be available at the time of the plenary).

1.2.2 Data collection

Semi-structured one-to-one cognitive interviews were undertaken with members of the public and carers (interviews with patients and social care users are on-going at the time of writing). Written informed consent was taken at the start of each interview. Participants completed a short survey (age, gender, ethnicity, any health condition they suffer from, any caring role they have, and EQ-5D-5L), though these questions were not compulsory. At the end of the interview, participants were compensated. All interviews were audio-recorded using an encrypted device and researchers also made brief notes. Ethical approval was obtained from the Institutional Review Boards and relevant Ethics Committees.

Each item was reviewed in turn and grouped by domain. Each participant saw only a subset of the domains resulting in them being asked about 30-40 items. Items were shown in a questionnaire format (Figure 3). Response options considered were frequency, severity, difficulty or agree-disagree.

Figure 3: Example questionnaire format

For each of the following statements, please tick one box that best describes your thoughts, feelings and activities over the **last 7 days**

	None of the time	Only occasionally	Some of the time / Sometimes	Often	Most or all of the time
I enjoyed what I did	☐	☐	☐	☐	☐

	No difficulty	Slightly difficulty	Some difficulty	A lot of difficulty	Unable
Because of hearing and/or speech, how difficult did you find it to have a conversation?					

A topic guide was used to support the interviews in the different countries. For each item, participants were asked to say how they would interpret the question and whether they would be able to answer it. Some items were similar with slightly different wording or framed in an either positive or negative way. In these cases, respondents were also asked whether they had a preference. Respondents were also asked for alternative wording where they highlighted problems with proposed wording. In some cases, different response options could apply and respondents were asked if they had a preference.

1.2.3 Data analysis

The feedback from the consultation groups was summarized and used to refine the questions before the face validity interviews. This included dropping questions that were problematic and adding or making changes to the existing questions to address the feedback.

The data generated from the interviews was analysed systematically by considering and documenting all feedback/comments reported by the respondents. Audio files were not transcribed verbatim, but were used alongside any notes made for accurate reporting. Using a spreadsheet, the researcher that conducted the interview made notes for each item that was discussed within the interview. This included noting the meaning/interpretation of the item, any positive or negative points raised with regards to an item and any alternatives suggested by the respondent. Where items were similar, preferred options were highlighted. Comments were paraphrased, unless they are perceived to be of sufficient value to be included verbatim in any written papers/reports, in which case they were flagged as a directly spoken response using quotes ("…"). This information was used to provide summaries for the items by domain. Information from the self-complete form was used to gain an understanding of the demographics of the sample, plus provide an opportunity to see whether particular issues with items arise more in certain groups than others.

E-QALY Sheffield team meetings were held after each subsequent 10 interviews. This allowed choice of items to be put forward for future face validity interviews to be modified depending on the findings of the earlier interviews, and the opportunity for some items to be discussed more where this may be beneficial, or discussed within particular demographic groups.

1.2.4 Training of interviewers

All interviewers, including in the international teams, were provided with training documents and videos and a topic guide to ensure that they undertook interviews to a similar level of quality (all materials available from the authors upon request). Primary investigators in each country were responsible for ensuring that interviews were undertaken in the same way.

2 Results

2.1 Items

The number of items generated per domain depended in part on the number of sub-domains. The number and wording of items also varied as the consultation process progressed. The 'activity' and 'feelings' domains had the most items (n=24 to 28), followed by 'relationships' (n=16), 'self-identity' (n=10), 'physical sensations' (n=8), and 'autonomy' and 'cognition' (n=7). The items are presented in Table 4 for 'activities' and in full in Appendix 1.

Items were presented in the face validity interviews by sub-domain e.g. the first six items in the activities domain (Table 4) related to 'Enjoyable or meaningful activity/role' sub-domain with some duplicates to assess framing e.g. "I did what I wanted to do" and "I could do the things I wanted to do" (Table 4). The aim was to identify which of the two options was preferred. Based on some initial feedback, we considered inclusion of aids in questions related to vision and hearing. We also considered inclusion of 'help received' as this was an important issue in the context of social care. Highlighted items at the end of the table were those that were not taken forward to face validation.

2.2 Consultation and Face validation

Face validation is still on-going, particularly with service users. We provide an example of interim findings based on the completed consultation with the PI group, members of NICE Citizens Council and the advisory group and some initial face validation interviews with the general population and carers using the activities domain. Table 4 presents items that were identified by the different groups as problematic and list some of the comments made. (Note that only a sub-set of those interviewed so far will have seen the activity items).

The consultation exercise around item generation and selection allowed both the selected items, and the selection criteria to be subject to scrutiny. Feedback from the consultation frequently focused on adherence/consistency of application of the selection criteria. The consultation exercise resulted in dropping (or modifying) some items prior to the face validity interviews. For the activities domain this resulted in dropping four potential items (shown at the bottom of Table 4) that were flagged in consultation as being complex and difficult to understand e.g. the term 'presentable', or being related to more than one concept, or where ideal states were included e.g. "I felt as clean and presentable as I wanted', or relating to social desirability issues e.g. someone being cared for may find it difficult to respond honestly to whether their 'home was as clean and as comfortable as they wanted'. Other items for which some concerns were raised during the consultation were taken forward to face validity with a view to generating more evidence on their suitability although concerns attached to these items were also taken forward.

The face validity interviews with carers and the general public identified some activity items as being ambiguous. This was either due to brevity and lack of context e.g. when questions referred to things people 'did', some participants wanted more information related to context. Some respondents in the face validity interviews noted potential ambiguity in the question, but that they would still find it fairly easy to answer the question.

Questions which referred to what individuals 'wanted' to do versus 'needed' to do were interpreted as expected with the former referring to what was preferred and the latter to activities that were essential such as activities of daily living.

Some items were interpreted in different ways e.g. 'communicate' referred to all kinds of communication –

telephone, conversation, text and email for one participant and another considered the item 'How well did you communicate with others?' to be assessing their skill in getting their message across effectively, another considered the response of others e.g. clinical staff not listening to them. This does not link to the original construct of hearing and speaking and points to ambiguity as to what answers would be referring to. Similarly, questions relating to difficulties with self-care were seen as arising from both physical limitations and resource limitations (e.g. lack of time).

The relevance of some items was also highlighted. This included comments around what could be reasonably expected e.g. 'everyone experiences boredom' or 'unrealistic to expect people to be able to do what they want'. There were also issues with questions related to self-care and receiving help for some carers who did not know why they would be asked these questions.

There were also comments in the consultation and face validity interviews in relation to framing – either the framing was not common usage, did not translate well or was not appropriate for valuation. Framing also referred to where the response options appeared e.g. at the end of a sentence was preferred to the middle for one advisory group member based on their experience. Where the main instructions appeared also made a difference e.g. some respondents forgot that the recall period was seven days when questions appeared in a tabular format with instructions at the top.

Table 4: Activity domain feedback (√ - flagged as problematic)

	Focus groups PI group (n=7) and NICE Citizens Council (n=5)	On-line survey Advisory group (n= 28)	Face validity interviews General public and carers (n= 13)	Comments to date
1 I enjoyed what I did		√	√	• Usual activities [that I did] may not be enjoyable • Ambiguous – not sure what it is referring to • Interpreted as encompassing a broad range of activities (dancing, gardening, meeting friends, playing scrabble, going out for lunch, shopping, watching TV)
2 I was able to do the things I value	-	-	√	• Ambiguous – too broad • What about adaptation. Wouldn't include not working in this response since that is now a permanent state. • Some would include household (e.g. cleaning) and caring tasks others would not.
3 I did things I found rewarding	√	√	√	• "rewarding" – complicates it • Rewarding for who – it may be rewarding for the caree and judged as important by the carer
4 I was bored		√		• Relevance – everyone experiences boredom
5 I did what I <u>wanted</u> to do	√	√		• Ambiguous – hard to get what is being asked • "unrealistic" – no one is expected to do anything they want
6 I could do the things I wanted to do				
7 I did what I <u>needed</u> to do	√	√		• "loaded question" for carer • Ambiguous – hard to get what is being asked. May feel 'unable' to fulfil caring role but have to get on with it anyway. • Doing things that are 'needed' is negative [carer] • Needed to do includes basic own self-care and caring tasks (relating to self-care others, food, hospital appointments)
8 I was able to do what I needed				
9 I had **no difficulty** with my day to day activities/ daily activities (e.g. working, shopping, travelling)		√	√	• Day-to-day activities not limited to these (also includes cleaning, having a shower, eating, emails, admin, gardening) • 'Difficulty' includes having the time and resources in addition to physical ability

10 Given the help I had/received my personal needs were met (e.g. being washed, going to the toilet, getting dressed, having food when I needed)	-	-	√	• 'received' preferred to 'had' • Not relevant - carer
11 Given the help I had/received my self-care needs were met (e.g. being washed, going to the toilet, getting dressed, having food when I needed)	-	-	√	• Ambiguous who this relates to [carer] • What if you didn't have help? • Not relevant - carer
12 I was able to look after myself (e.g. being washed, going to the toilet, getting dressed, having food when I needed)		√		• Carers may be capable of looking after themselves but not have the time
13 I needed help with looking after myself (e.g. being washed, going to the toilet, getting dressed, having food when I needed)	-	-		
14 I was able to look after myself with **no difficulty** (e.g. washing, dressing, going to the toilet)		√		
15 I had **no difficulty** with self-care activities (e.g. washing, dressing, going to the toilet)				
16 I was able to get around inside my home with **no difficulty**		√		
17 I was able to get around outside with **no difficulty**		√	√	• What about aids? What about driving? • Ambiguous – needs examples • linked to one above – do you need both?
18 How well did you communicate with others?	√	√	√	• Translation of this structure "How…" into Latin languages may be difficult • Ambiguous what communicate means (talk, listen, internet, response of others?) • Some interpret as an ability to get one's point across effectively (e.g. talking to health professionals) • Framing difficult for valuation
19 I was able to communicate with others with **no difficulty**				• Ambiguous what communicate means (talk, listen, internet, response of others?)
20 Because of hearing and/or speech, how difficult did you find it to have a conversation?		√		• Not easy to read/understand
21 How well can you hear (using hearing aids if needed)?				
22 I had **no difficulty** hearing (using hearing aids if needed)				
23 How well can you see (using your glasses or				• Framing difficult for valuation

contact lenses if they are needed)?				
24 I had **no difficulty** seeing (using your glasses or contact lenses if they are needed)				
I felt as clean and presentable as I wanted	√	√	-	• 'Presentable' – difficult word • Social desirability issues – value-laden and intrusive • 'as I wanted' – could be room for improvement even for those without a problem • double-barrelled • who defines 'clean'
My home was as clean and comfortable as I liked	√	√	-	• difficult for a '9 year old' • Social desirability issues – judgemental about care received • double-barrelled
I get all the food and drink I like when I want?	√	√	-	• Ambiguous – not clear what is being asked
I had [insert response] doing everyday activities (e.g. washing, dressing, going to the toilet)	√	√	-	• 'everyday activities' is broader than examples given • May be value laden – not everyone will wash every day

Note, advisory group members could flag the item without commenting

3 Discussion

The E-QALY project aims to develop a broader generic measure of QoL for use in economic evaluation to meet a perceived need amongst some decision makers. Methods of development draw upon current good practice for measure development.

Development began with a conceptual framework that is an extended version of the Wilson and Cleary model for health to incorporate a broader range of outcomes for health care, social care and carers. A review of qualitative literature on quality of life was undertaken in Stage 1 of the project to provide the initial domains and sub-domains for the measure. Stage 2 was a generation of items based on terms from the qualitative review and items from existing health and wellbeing measures. These items were modified and an initial selection undertaken by the research team to ensure coverage of the concepts underlying the domains and sub-domains and to meet selection criteria of good item construction. Further modification of items occurred following consultation with key stakeholder groups from the project PI group, members of NICE Citizens Council and the advisory group. About 100 items were taken forward to face validity interviews, of which 14 had been conducted in the UK at the time of writing. A large number of items were taken forward into the face validity work. This is intentional to ensure the face validity and psychometric analysis can meaningful inform the selection of the best items and ensure all key concepts are covered.

The generation, selection and testing of items for the new measure is a large logistical exercise and one that involves: (1) many stakeholders (patients, service users and decision makers) and (2) is truly an international endeavour by ensuring the beneficiaries across 6 countries have a say.

Initial findings, from the consultation exercise and early face validity interviews (as shown in the activities domain example findings) flags a lot of items as problematic. Even previously published items are flagged as problematic when subject to face validity interviews across the different groups. Short items without additional context raise concerns and uncertainties about their scope yet longer items risk problems with readability. Identifying items that work well and share a consistent meaning across these different groups, (different types of patients, the general public, social care users and carers) is likely to be challenging.

The importance of conducting face validity interviews across different groups is evidenced in the varied interpretations arising from the same items. For example, being able to communicate well from a patient perspective has a physical emphasis, for some non-patients/carers this is interpreted as how successfully they reveal communication skills.

These are preliminary results – we await results from face validation with patients and other service users, and the results of similar work in other countries – much of which we hope to be available for discussion at the plenary meeting in September.

References

Al-Janabi H, Flynn T, Coast J. Estimation of a preference-based Carer Experience Scale. Medical Decision Making. 2011; 31(3):458-68.

Bradburn NM, Sudman S, Wansink B. Asking questions: the definitive guide to questionnaire design--for market research, political polls, and social and health questionnaires. John Wiley & Sons; 2004 May 17.

Brazier JE, Connell J, Papaioannou D, Mukuria C, Mulhern B, O'Cathain A, Barkham M, Knapp M, Byford S, Gilbody S, Parry G. 2014. Validating generic preference-based measures of health in mental health populations and estimating mapping functions for widely used specific measures. Health Technology Assessment; 18(34).

Brouwer WB, Van Exel NJ, Van Gorp B, Redekop WK: The CarerQol instrument: a new instrument to measure care-related quality of life of informal caregivers for use in economic evaluations. *Qual Life Res* 2006,15(6):1005–1021.

Butt Z, Reeve B. Enhancing the patient's voice: Standards in the design and selection of patient-reported outcomes measures (PROMs) for use in patient-centered outcomes research. Contracted report for the Patient-Centered Outcomes Research Institute (PCORI). 2012 Mar 30.

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):874-882.

Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. John Wiley & Sons; 2016. Third Edition

Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. The European Journal of Health Economics. 2017 May 30:1-4.

Jobe JB. Cognitive psychology and self-reports: Models and methods. Quality of Life Research. 2003 May 1;12(3):219-27.

Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ open. 2016 Jul 1;6(7):e010641.

Longworth L, Yang Y, Young T, Hernandez Alva M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P, Keetharuth A, Brazier J. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: systematic review, statistical modelling and survey. 2014.

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. Second edition. Oxford University Press, New York. 1996.

Netemeyer RG, Bearden WO, Sharma S. Scaling procedures: Issues and applications. Sage Publications; 2003 Mar 12.

Netten A, Burge P, Malley J, Potoglou D, Towers AM, Brazier J, Flynn T, Forder J. Outcomes of social care for adults: developing a preference-weighted measure. Health Technology Assessment. 2012;16(16):1-66.

Peasgood T, Carlton J, Brazier J. A qualitative study of the views of health and social care decision makers on the role of wellbeing in resource allocation decisions in the UK, 2016.

Richardson J, McKie J, Bariola E. Multiattribute utility instruments and their use. In: Culyer AJ (ed). Encyclopaedia of Health Economics, vol.2 San Diego: Elsevier; 2014:341-357.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 2008.

Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, Parkinson J, Secker J, Stewart-Brown S. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health and Quality of life Outcomes. 2007 Dec;5(1):63.

Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. Jama. 1995 Jan 4;273(1):59-65.

Wisløff T, Hagen G, Hamidi V, Movik E, Klemp M, Olsen JA. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. Pharmacoeconomics. 2014 Apr 1;32(4):367-75.

Appendix 1: Proposed items

Autonomy	Cognition	Feelings and emotions	Physical sensations	Relationships	Self-identity
1 I felt able to cope	1 I found it hard to concentrate	1 I felt happy	1 I had no pain…	1 I felt supported by other people	1 I felt confident in myself
2 I felt unable to cope	2 I found it hard to focus my thoughts	2 I felt unhappy	2 How often do you experience pain	2 I felt unsupported	2 I felt confident
3 I felt unable to cope with my day to day life	3 I found it hard to pay attention	3 I felt depressed	3 I had **no** discomfort e.g. feeling like throwing up, breathless, itching etc. (but not including pain)	3 Other people gave me support	3 I felt I was treated with respect
4 I felt overwhelmed by my problems	4 I had trouble thinking clearly	4 I felt sad	4 I had discomfort e.g. feeling sick, breathless, itching etc. (but not including pain)	4 I had support when I needed it	4 I felt respected
5 I felt in control of my daily life	5 I had trouble remembering	5 I enjoyed life	5 I felt exhausted	5 I had disagreements and conflict with people	5 I felt like I lived with dignity
6 I felt in control of my day to day life	6 I had trouble with my memory	6 I felt content with my life	6 I got tired easily	6 I got on with people around me	6 I felt unsure about myself
7. Which of the following	7 I felt/was confused	7 I thought my life was not	7 I was too tired to do anything	7 I got along well with people	7 I felt good about myself

Autonomy	Cognition	Feelings and emotions	Physical sensations	Relationships	Self-identity
statements best describes how much control you have over your daily life? …		worth living		I came into contact with	
	I had trouble making decisions	8 I felt that I had nothing to look forward to	8 I had problems with my sleep	8 I felt lonely	8 I felt like a failure
	I was able to make decisions	9 I had nothing to look forward to		9 I feel there was nobody I was close to	9 I felt valued
		10 I looked forward to each day		10 I felt I had no one to talk to	10 I felt useful
		11 I felt frightened		11 I felt isolated	I felt positive about myself
		12 I felt afraid		12 I felt people avoided me	
		13 I felt safe		13 I felt judged by others	
		14 I felt unsafe		14 I felt accepted by others	
		15 I felt secure		15 I felt excluded	
		16 I felt anxious		16 I felt left out	
		17 My worries overwhelmed me		I felt close to others	
		18 I felt worried		I felt humiliated	
		19 I felt calm		I felt I was a burden to others	
		20 I felt relaxed		I had to rely on others to take care of me	
		21 I felt irritable			
		22 I felt irritated			

Autonomy	Cognition	Feelings and emotions	Physical sensations	Relationships	Self-identity
		23 I felt angry			
		24 I felt frustrated			
		25 I lost my temper easily			
		I felt cross			

Items in grey were not taken forward to face validation