

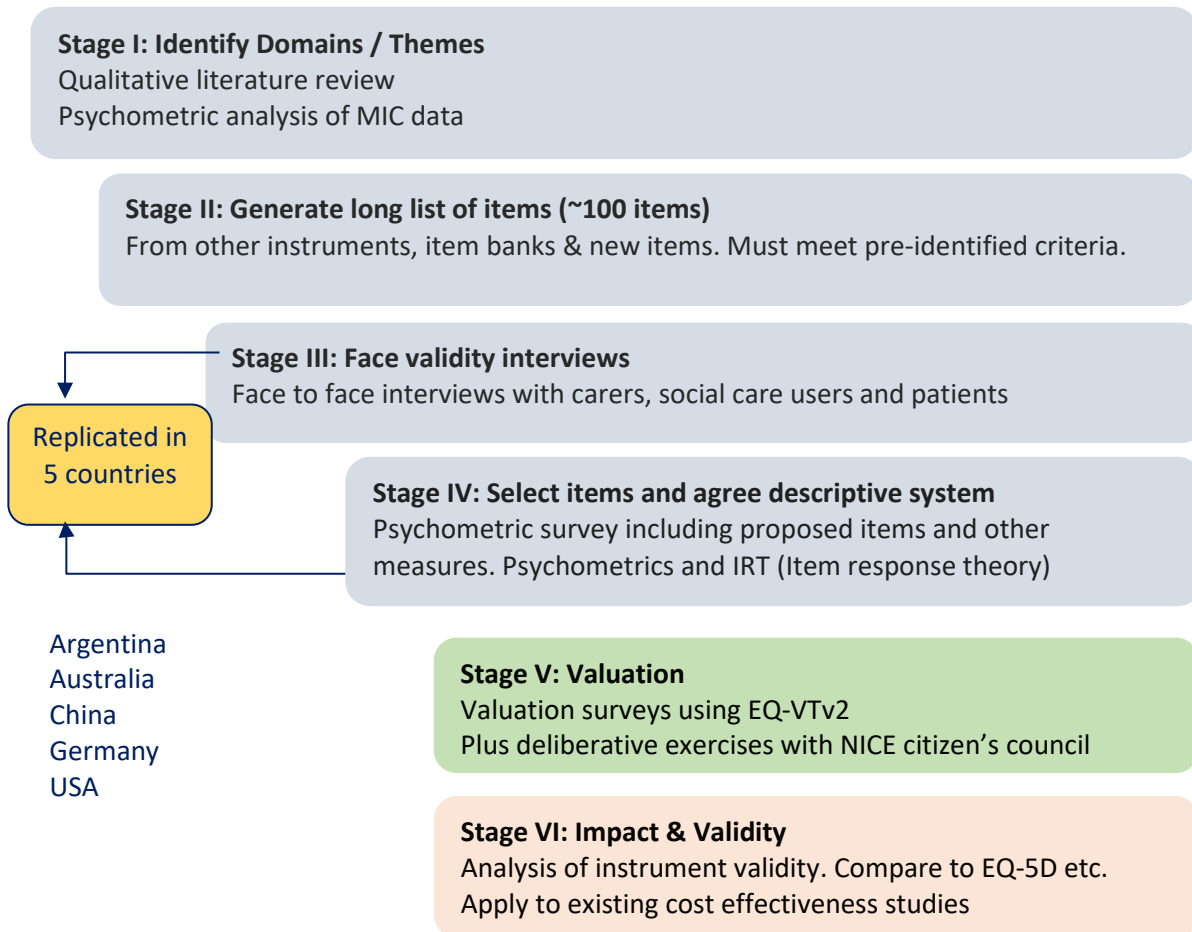
Extending the QALY: Psychometric results summary (Stage 4) October 2019

1 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 (care for people with a disability and care for the elderly) and public health, based on the views of users and beneficiaries of these services.

The project has six stages to identify the potential domains, generate a long list of items (or questions), identify suitable items from the long list based on face validity and psychometric analysis, valuation and validation of the new measure (Figure 1). The aim is to develop a long questionnaire and a shorter classification system the latter which will be used for valuation on a 0 to 1 QALY scale.

Figure1 : Extending the QALY stages



Stages 1 to 3 identified the domains and sub-domains for the new measure, identified questions and tested the questions in semi-structured interviews. Selected questions from stage 3 were then tested in large quantitative surveys with patients, social care users, carers and members of the general population in Stage 4, the psychometric analysis stage. Results from this stage and from stage 3 were used to make overall recommendations of potential questions to include in a longer measure. This report provides an overview of the psychometric analysis methods and a summary of the findings.

2 Methods

2.1 Sample and data collection

Table 1 summarises the aim, sample and data collection for psychometric analysis. The aim was to test the performance of the items quantitatively to assess whether items were valid in patients, social care users and carers. Data was collected in these populations drawn from the United Kingdom (UK), Argentina, Australia, China, Germany and the USA. Ethical approval was obtained from the Institutional Review Boards and relevant Ethics Committees.

Table 1: Psychometric analysis methods summary

Aim	<p>To assess:</p> <ul style="list-style-type: none"> - Whether questions are relevant to participants based on distribution of responses and level of missing data - Whether items are able to distinguish between known groups (groups for whom we would expect to answer differently e.g. those with a health condition) - Whether the items fit in the expected sub-domain or domain construct - Whether items perform well in terms of distinguishing this construct
Sample	<ul style="list-style-type: none"> - Patients, social care users, carers and members of the general population - Participants from each group in United Kingdom (UK) but different groups in Argentina, Australia, China, Germany and the USA - Sample size target: n= 2000 UK; n=500 all other countries - Age 18 and above, capacity to consent - Participants were compensated with vouchers or online points
Data collection	<ul style="list-style-type: none"> - Paper (sub-sample in UK only) or online (all countries including UK) - Self-complete (option for interviewer-administered in UK but this was not requested) - E-QALY test questions formatted as a questionnaire alongside EQ-5D-5L, EQ-5D-3L, a mental wellbeing measure (Short Warwick Edinburgh Mental Wellbeing Scale) and a social care measure (ASCOT). - Also completed demographic questions - Data collection managed by a single company in England. - Translated items from Stage 3 were used; where new or amended items were proposed, international teams proposed translations. All items were then subjected to a back translation to English to ensure accuracy

2.2 Analysis

The data was used for a number of complementary analysis. An analysis protocol was used to support consistent analysis across the countries with modification based on sample size and groups that were included. The terms in [red](#) link to table headings within the findings which are provided separately in a zip file, disaggregated by country.

a. Distribution of responses

We examined the responses to each item based on missing data (in the UK paper version only) and the distribution of responses. Items with a spread of responses across all the responses choice should help us distinguish between different levels of quality of life.

Criteria for assessment of distributions:

We flagged items as potentially problematic if they had a really skewed **distribution** such that either more than 70% or less than 5% responded in the top or bottom category. Questions with a very skewed distribution were penalised if they were trying to pick up a construct which we would expect to be more evenly distributed in the sample. This was judged in part in relation to other items tapping into the same construct, and in part our own understanding of the construct. For example, questions asking about vision, hearing, self-care and mobility are expected to have a skewed distribution.

b. Sensitivity of items to known group difference

We expected the answers from people with health conditions to show poorer quality of life than those people without health conditions. This should be reflected in a poorer response choice of people in a particular 'known group' such as patients with arthritis which is measured as an '**effect size**'. We assessed the ability of questions to identify known group differences based on physical health conditions and mental health conditions compared to no condition (matched for age and gender). Known group differences based on being a carer (matched for age, gender and health) and number of hours cared were also assessed.

Criteria for assessment of know group differences:

We judged items using standard cut offs for effect sizes [0.2 to <0.5 = small 0.5 to < 0.8 = medium, >=0.8 large effect sizes]. We expected effect sizes to be negative indicating poor health. Questions with low effect sizes were penalised. We had no strong prior beliefs about whether carers should give different answers to many questions – so we only penalise questions which have a low effect size for the carer group when other similar questions show a higher effect size.

c. Test of dimensionality / number of separate constructs the questions can identify

We assessed the dimensionality based on the overall survey responses; this shows us the number of separate constructs the data can identify. We intended the questions in the survey to cover the constructs (or dimensions) from our conceptual model arising from Stage 1 of the project. This step of the analysis helps to confirm that the items are picking up what we expected them to.

We looked at correlations between the items, and methods to explore the patterns within the data including exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA results were used to assess whether it was necessary to take into account whether questions were positively or negatively framed (i.e. asking about happiness rather than unhappiness). The CFA model was used to confirm the conceptual model while relying on the EFA and correlations to inform the process. Dimensionality was confirmed using UK data and then the resulting models was tested in the five other country data sets with modification to achieve model performance criteria.

Criteria for assessment of the influence of a positive or negative measurement factor (MF):

We found that the answers to many questions were influenced not just by the construct or domain (such as mobility, coping or safety) but also by whether the question was asked in a positive way or in a negative way. People tend to treat positively framed questions (such as feeling calm), slightly differently to negatively framed questions (such as feeling anxious).

We penalised questions where the responses were being influenced more by this 'positive wording' or 'negative wording' measurement issue. We flagged items if the ratio of the loading to factor of interest (a measure of association between the item and the construct we are trying to measure) was below 1.75 times the loading onto the measurement factor.

d. Tests using Item Response Theory (IRT)

We used the questions within each construct (or domain) to estimate a new variable or latent factor that is a prediction of how well an individual respondent would score on an unobserved scale of the construct of

interest (e.g. pain, coping, safety). We explored how good each question was at providing information about the latent score for each respondent. An estimated higher level of the latent pain score, for example, should mean the individual is more likely to answer using a higher frequency response option (i.e. they experience pain more often).

We used item response theory (IRT) to produce graphs showing the estimated probability of answering in each category (known as category characteristic curves). If we knew the level of a construct experienced by an individual then an item for which we can accurately predicted which response option they would choose is preferred.

Criteria for assessment of the 'order' of response choices:

We penalised items where the response choices were not clearly ordered into 5 categories.

Criteria of assessment for providing information on the construct:

Item Information Functions are another type of graph which can be produced from IRT analysis showing how well and how accurately each question measures the latent factor at various levels of the latent factor, using a measure called 'information'.

We penalised questions if they showed much less information than similar questions, particularly if they did not provide information over a wide range of the latent score, especially on the poorer end of health/wellbeing. To assess this we studied the **slope** (a higher value shows greater ability to distinguish between levels of the contrast), the **thresholds** (the cut offs between predicting the most likely response option and the next highest – because the items are coded such that a higher number represents worse quality of life we prefer items which the thresholds cover the positive range of the latent score), and the **range** (the range of the latent construct in which the item provides information, where a wider range is preferable in this case unless we are able to include more than one item about the construct in the final measure).

Criteria of assessment for consistent performance of items between groups:

We are also interested in whether different groups treat the questions in the same way. We looked at five groups: age (<45 or 45+), gender, education (having a degree or not), having a mental health condition or not, and being a carer or not - to see if being in one of these groups influenced respondent answers when we controlled for their predicted level of the latent construct. We want a question about 'coping', for example, just to be about the levels of coping someone feels – if we find that women and men with the exact same level of the latent score for coping answer the questions tapping into that construct in a different way this isn't ideal. This is known as Differential Item Functioning (DIF). We penalise questions which show **DIF**, although we do not over penalise them since it is quite hard to find questions about health and wellbeing that do not show any DIF.

Additional considerations

In making the overall judgements on item performance we also considered whether the item fitted comfortably into the overall construct used in the IRT. This is assessed by a measure called item fit which is assessed by the statistic **S- χ^2** where this is significant (we use a p value of <0.01 because we are running multiple tests) this suggests the item is measuring something slightly different to the other items in the construct. Whether this is used to penalize the item is judged on a case-by-case basis depending on the possible reasons for misfit e.g. it could be because its positively worded and other items are negatively worded, it could be that the item was intended to pick up something slightly different, or it could be an indication that the item is not performing as expected.

We also considered whether any items were linked in ways that were not explained by their correlation to the latent construct. The IRT models rely on the assumption that this is not the case, and it is usual to test for local dependency (**LD**) before relying upon results of IRT analysis. During instrument development items with problems with ordering, local dependence or DIF may be excluded or amended (e.g. response options

combined where ordering problems are identified); in this case all items are retained, mostly to maintain consistency in analysis across 6 countries, however, where problems have been identified we interpret the IRT findings with a lighter touch.

Each country team assessed their evidence and summarised it based on the criteria. An overall summary across the psychometric evidence was given where items were scored on a 4 to 1 scale where:

4 - item performs very well

3 - item performs fairly well

2 –item performs weakly or there is mixed evidence on item performance

1 - item performs poorly - include the reasons as to why the item is placed in this category e.g. not culturally relevant

This was combined with the face validity results in order to provide an overall assessment of the evidence for the questions.

3 Results

3.1 Sample

Table 3 shows the mix of participants that were recruited to take part in the psychometric survey across the six countries. Most countries recruited across two or more groups. Patients were recruited across those with different conditions drawn from physical health and mental health.

Table 3: Psychometric survey participants

	Long-term condition	Social-care users	Carers	Mean (SD) Age	Female (%)	Total
United Kingdom	906	226	393	48.6 (18.8)	57	1923
Australia	374	-	115	49.9 (16.9)	39	514
Argentina	317	287	339	37.4 (12.8)	40	497
China	357	-	226	41.4 (15.5)	60	497
Germany	333	131	280	44.8 (17.1)	51	496
USA	623	86	196	53.8 (17.5)	48	903

3.2 Missing items, distribution and known group validity

The proportion of missing data ranged from 0.6 to 4.5% (UK paper survey) with the exception of one question (able to cope) which was 19.5% but this was due to a problem with the printing of one version of the questionnaire. Most of the items did not have high ceiling effects with the exception of items in the self-care, mobility, hearing and seeing sub-domains. Items did not have <5% reporting at the highest end, except 'I felt worried' in Argentina, and 'I felt able to cope with my day-to-day life' in China.

There was evidence of less than 5% at the lowest level for most items, with exception of the UK data.

Most of the items were able to detect known group differences across the physical and mental health conditions with moderate to large effect sizes. Some of the items were able to distinguish between carers, but there were small effect sizes when carers were compared to non-carers and when carers with who spent a low number of hours caring were compared with those who spent a high number of hours spent caring.

3.3 Domain structure

Exploratory Factor Analysis (EFA) indicated that the positive and negatively worded questions had a 'measurement' factor that was separate to the construct of interest. The CFA included positive and negative factors alongside the domains drawn from the underlying conceptual model. Most of the variance was explained by the constructs/domain factors not the measurement factors.

The conceptual model was reasonably well confirmed, particularly across UK, Australia, USA, Argentina and Germany, with some modification in the latter two countries (Figure 1). The relationship sub-domains were combined as were the cognition sub-domains, plus the sub-domain of Daily Activities was not well defined and items were combined with the 'Mobility' and 'Meaningful/enjoyable activity' sub-domains. One item (*I was able to do the things I wanted to do...*) originally intended to measure the 'Meaningful/enjoyable activity' subdomain fitted best within the 'Control' domain. However, this may have been due to ordering effects with the survey as this item appeared next to an item asking about control.

Modifications for Argentina and Germany included combining mobility and self-care, removing energy and combining self-worth and coping for Argentina. The model did not fit the China data as well; many of the feelings sub-domains needed to be combined. There was evidence of high correlation between factors but combining them did not improve model fit.

3.4 IRT results

Results from the IRT analysis indicate that most of the items were ordered. The items '*I had support when I needed it*', '*I thought my life was not worth living*' and '*I felt like a failure*' were disordered in some countries. This problem was due to a lack of distinction between levels 1 and 2 and 4 and 5 which may indicate that less levels are required for the response options for these items. There was evidence of DIF for many of the items.

Items which did not fit well in the IRT model or were relatively strongly influenced by the positive/negative measurement factors in the CFA often perform poorly within the IRT analysis. This is interpreted in the light of whether they are conceptually aiming to measure something slightly different to the other items within the construct.

3.5 Overall review of the evidence

Table 3 shows the summary review of the psychometric evidence and face validity as well as an overall review of the evidence by country.

4 Discussion

4.1 Overall findings

Based on the level of missing data and distribution across response levels, the questions were relevant. However, missing data was only tested in the UK where paper-based questionnaires were completed. There was some evidence of skewed responses but this was mostly in areas where skewed data was expected.

Majority of the items were able to distinguish between those with physical and mental health conditions as well as by severity where this was tested. However, the evidence was mixed for carers. Caring can both improve and decrease wellbeing and this may have impacted on the results.

The conceptual model was generally confirmed although there was evidence of high correlation between factors. There was a positive and negative measurement factor along with the construct/domain factors

but the latter tended to explain a larger proportion of the variance of responses. IRT results indicated that most items performed well apart from in the context of DIF.

4.2 Strengths and Limitations

The psychometric analysis relied on large mixed samples in different countries which allowed multiple testing to be undertaken. This included testing in non-English-speaking contexts. However, this was also a limitation as it made it more difficult to undertake in-depth analysis of each dataset e.g. IRT models were not tested by excluding items with local dependency. Further testing is recommended.

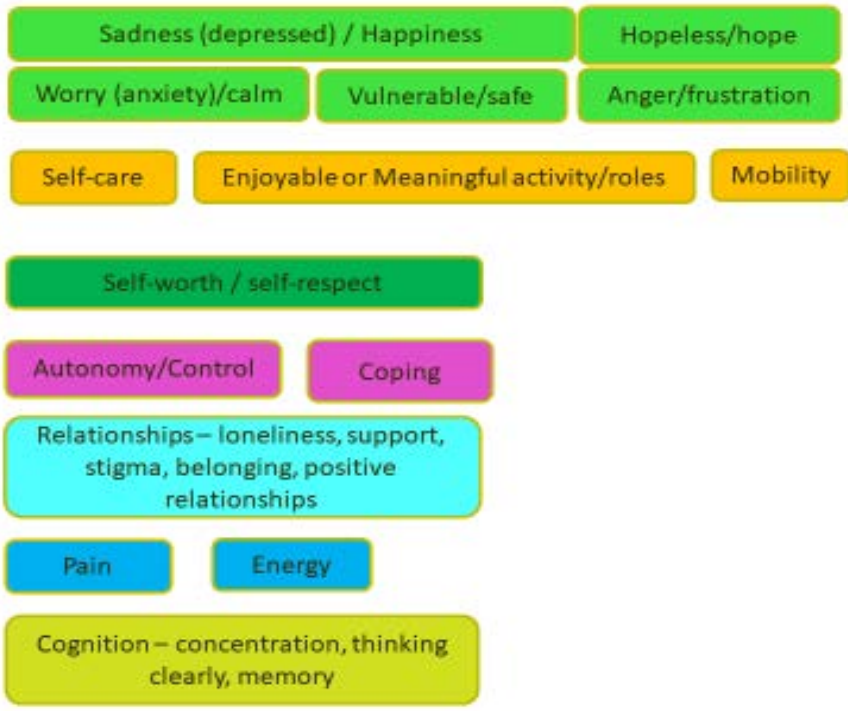
A mix of patients with physical and mental health conditions as well as social care users and carers were targeted in the different countries which enables assessment of the questions in future users. This allowed assessment of known group differences. However, there was limited assessment based on severity and no assessment of change over time which is an important aspect of the performance of measures; further validation in this context is therefore needed.

Using online samples allowed different condition groups to be targeted in a relative short and inexpensive way compared to collecting data using different recruitment strategies. However, this meant that some assessment such as missing data could not be undertaken. This was partly mitigated in the UK where both online and paper based completion was possible but those online were younger than those who completed the measure by paper.

4.3 Conclusion

The results from the psychometric analysis will inform which questions are taken forward for a long questionnaire and a shorter version which will be valued on the 0 to 1 scale.

Figure 1: Confirmatory Factor Analysis for UK, Australia and USA data and model fit statistics for all countries [excluding the domains Seeing, Hearing, Sleep problems and Discomfort]

	RMSEA (95% CI)	CFI	TLI	
	UK: (full model) 0.059 (0.059 0.060)	0.958	0.953	
	Argentina: Model not positive definite correlation matrix until <i>Mobility</i> merged with <i>Self-care</i> , <i>Energy</i> removed and <i>Self-worth</i> merged with <i>Coping</i>	0.049 (0.047 0.051)	0.956	0.952
	Australia: (full model) 0.057 (0.055 0.059)	0.965	0.961	
	China: Model not positive definite correlation matrix until <i>Mobility</i> merged with <i>Self-care</i> <i>Anxiety</i> , <i>Happiness</i> , <i>Hope</i> , <i>Coping</i> , <i>Self-worth</i> all combined to one factor <i>Energy</i> removed, Some items did not fit either the positive or the negative measurement factor	0.072 (0.070 0.074)	0.907	0.901
	Germany: Model not positive definite correlation matrix until <i>Mobility</i> merged with <i>Self-care</i> , and <i>energy</i> removed.	0.052 (0.050 0.054)	0.961	0.957
	USA (full model) 0.055 (0.054 0.057)	0.969	0.966	

Each box represents a factor from the CFA. Similar colours represent sub-domains from the same theme. RMSEA (Root mean square error) < 0.6 taken as good; CFI (comparative fit index) > 0.95 taken as good, > 0.90 taken as adequate; TLI (Tucker-Lewis index) > 0.95 taken as good, > 0.90 taken as adequate

Sadness/Happiness: happy, unhappy, cheerful, sad; **Hopeless/hope:** hopeful, looking forward, life not worth living, swemwbs optimistic; **Worry/calm:** calm, worried, anxious, swemwbs relaxed; **Vulnerable/safe:** safe, unsafe, afraid, frightened; **Anger/frustration:** angry, frustrated, irritable, lost temper; **Self-care:** difficulty meeting personal needs, personal needs met, able to look after self, difficulty washing etc.; **Enjoyable roles/activity:** enjoyed what I did, did things I valued, did things I needed, did thing I wanted; **Mobility:** get around inside, get around outside, able to do activities; **Self-worth:** confident, felt good about myself, felt unsure, felt like a failure; **Autonomy/Control:** control, no control, control with definition, swewmbs able to make up my mind; **Coping:** able to cope, unable to cope, overwhelmed by problems; **Relationships:** unsupported, support when needed, got along well, lonely, nobody close to, no one to talk to, avoided, accepted, excluded, left out, isolated; **Pain:** pain severity, pain frequency, eq-5d pain; **Energy:** very tired, exhausted; **Cognition:** concentrate, pay attention, thinking clearly, confused, trouble remembering.

Table 3: Overall review of the evidence based on face validity and psychometric results

Domain	Item	UK			Argentina			Australia			China			Germany			USA		
		FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR
Activity																			
Meaningful / enjoyable	I enjoyed what I did	3	3	3	4	3	3	4	4	4	2	3	3	4	3	3	4	4	4
	I was able to do the things I value	3	4	3	3	3	3	3	3	3	3	3	3	4	3	3	2	3	2
	I could do the things I wanted to do	3	4	3	4	3	3	1	3	2	3	2	2	4	3	4	3	3	3
	I was able to do the things I wanted to do (S1)	-	3	3	-	3	3	-	2	1	-	3	2	-	3	3	-	2	2
Self-care	My personal needs were met	3	1	1	2	2	2	3	1	1	3	2	2	2	1	1	4	1	1
	How well were your personal needs met (D)	-	3	3	-	2	3	-	3	3	-	4	4	-	3	3	-	2	2
	I was able to look after myself	3	1	1	4	2	2	3	1	1	3	2	2	3	1	2	4	1	1
	How difficult is it for you to wash, toilet, dress yourself, eat or care for your appearance? (D1)	-	3	3	-	2	2	-	3	3	-	3	3	-	2	2	-	3	3
Mobility	How well were you able to get around inside your home (D)	4	3	3	4	3	3	3	3	3	4	3	3	3	3	3	4	2	2
	How well were you able to get around outside your home (D)	3	3	3	4	4	4	3	4	4	4	4	4	2	3	3	4	3	3
	How well were you able to do your day to day activities (D)	-	3	3	-	4	3	-	3	3	-	3	3	-	3	3	-	1	2
Hearing and speech	Because of hearing and /or speech, how difficult did you find it to have a conversation? (D)	4	3	3	3	3	3	1	2	1	2	3	3	3	2	3	4	2	3
Hearing	How well can you hear (using hearing aids if you normally wear them)? (D)	3	3	3	3	3	3	3	2	2	2	3	2	3	3	3	4	3	3

Domain	Item	UK			Argentina			Australia			China			Germany			USA			
		FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	
Seeing	How well can you see (using your glasses or contact lenses if they are needed)? (D)	4	3	3	3	3	3	3	2	2	2	2	2	2	3	2	3	4	2	3
Relationships																				
Support	I felt unsupported by other people	3	4	3	3	3	3	2	3	3	4	3	3	1	3	2	2	3	3	3
	I had support when I needed it	4	1	1	4	3	2	3	2	2	4	3	3	3	2	3	3	3	2	2
Positive relations	I got along well with people around me	-	2	2	3	3	3	-	1	1	-	2	3	-	1	1	-	1	1	
Lonely	I felt lonely	4	4	4	4	3	3	3	4	4	4	3	4	3	3	3	3	3	3	3
	I felt there was nobody I was close to	2	4	3	4	3	3	1	3	2	4	4	4	1	3	2	2	4	3	
	I felt I had no one to talk to	4	3	3	3	3	3	1	3	2	3	4	4	3	3	3	4	3	3	
Stigma	I felt people avoided me	2	3	2	3	2	2	3	2	3	4	3	3	4	3	4	1	3	2	
	I felt accepted by others	4	3	3	4	3	3	3	3	3	4	3	3	4	2	3	4	3	3	
Belonging	I felt excluded	4	4	4	4	3	3	3	3	3	4	4	4	3	3	3	2	3	2	
	I felt left out	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	3	3	
	I felt isolated	3	4	3	2	2	2	3	3	3	4	4	4	2	3	2	1	3	2	
Cognition																				
Concentrate	I found it hard to concentrate	4	4	4	4	3	4	1	3	2	3	4	4	3	3	3	4	3	3	
	I found it hard to pay attention	4	4	4	3	3	3	3	4	4	4	3	3	3	4	4	4	2	2	
	I had trouble thinking clearly	4	4	4	3	3	3	1	3	2	4	4	4	2	3	2	1	3	2	
Memory	I had trouble remembering	3	4	3	4	3	3	1	3	2	3	4	3	4	4	4	3	3	3	
Confusion	I felt confused	2	4	3	2	3	1	3	2	2	2	3	3	1	3	2	1	3	2	

Domain	Item	UK			Argentina			Australia			China			Germany			USA		
		FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR
Self-identity																			
Confident	I felt confident in myself	3	3	3	4	2	3	3	3	3	3	3	3	2	2	2	4	2	3
	I felt unsure about myself	3	4	3	3	3	3	3	3	3	2	2	2	1	3	2	1	2	2
Self-worth	I felt good about myself	4	3	3	4	3	3	3	2	2	3	3	3	3	3	3	4	3	3
	I felt like a failure	2	3	2	4	3	3	3	2	2	4	3	2	1	3	2	1	1	1
Autonomy																			
Cope	I felt able to cope with my day to day life	3	3	3	4	3	3	3	2	3	3	2	2	4	4	4	4	2	3
	I felt unable to cope with my day to day life	4	3	3	2	3	3	1	3	2	3	3	3	3	3	3	2	3	2
	I felt overwhelmed by the problems or situation	-	4	4	-	3	3	-	3	2	-	3	2	-	3	3	-	3	2
Control	I felt in control of my day to day life	4	3	3	3	2	2	3	3	3	2	2	2	2	3	3	4	2	3
	I felt in control of my day to day life (By control we mean...)	-	4	4	4	3	3	-	2	2	-	4	3	-	3	3	-	2	2
	I felt I had no control over my day to day life	-	3	3	-	4	3	-	3	3	-	3	3	-	2	2	-	1	1
	I was able to do what I needed	4	3	3	4	3	4	1	2	1	3	2	2	3	2	3	3	2	3
Feelings																			
Happiness	I felt happy	3	3	3	4	2	3	3	3	3	3	3	3	4	2	3	4	3	3
	I felt cheerful	-	3	3	-	2	3	-	3	3	-	2	3	-	2	2	-	3	3
	I felt unhappy	3	3	3	2	2	1	3	2	3	2	4	3	3	3	3	2	2	2
	I felt sad	2	3	2	4	4	3	3	2	3	2	3	3	4	3	4	3	1	2
Hope	I thought my life was not worth living	3	2	3	4	4	4	1	3	1	2	4	3	1	3	2	2	2	2
	I felt that I had nothing to look forward to	2	3	3	4	4	4	3	3	3	1	3	2	2	4	3	2	3	3
	I felt hopeful about my future	-	3	3	-	3	3	-	1	1	-	3	2	-	2	2	-	2	2

Domain	Item	UK			Argentina			Australia			China			Germany			USA		
		FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR	FV	PV	OR
Safety	I felt frightened	2	3	2	1	3	2	3	3	3	4	3	4	3	3	3	1	3	2
	I felt afraid	2	3	2	4	3	3	3	2	3	4	4	4	3	3	3	4	3	3
	I felt safe	3	3	3	3	3	3	3	2	3	2	2	2	2	3	2	2	2	2
	I felt unsafe	3	4	3	2	4	2	3	2	3	2	4	3	1	3	2	2	2	2
Anxiety	I felt anxious	4	4	4	3	3	3	3	3	3	3	4	3	2	4	3	4	3	3
	I felt worried	4	3	3	4	3	3	3	3	3	3	3	3	3	4	4	4	4	4
	I felt calm	3	2	2	4	2	2	3	2	2	4	3	3	2	1	1	4	2	3
Anger	I felt irritable	4	3	3	3	4	4	3	3	3	4	3	3	3	2	3	4	2	3
	I felt angry	3	3	3	4	3	3	3	3	3	4	3	3	4	3	4	4	3	3
	I felt frustrated	4	3	3	4	3	4	3	3	3	4	2	3	4	2	3	4	3	3
	I lost my temper easily	4	2	3	4	2	3	1	3	2	4	3	3	2	2	2	1	3	2
Physical Pain	I had physical pain (S)	4	4	4	4	3	4	4	3	3	3	4	4	3	3	3	3	3	3
	How often did you experience pain	4	3	3	3	3	3	4	3	3	3	4	4	4	3	4	3	4	3
Discomfort	I had physical discomfort (S)	4	4	4	4	4	4	3	3	3	3	3	3	3	3	3	4	3	3
	I had physical discomfort	4	4	4	4	3	3	3	3	3	3	3	3	3	4	3	4	4	4
Energy	I felt exhausted	4	4	4	4	4	4	2	3	3	4	4	4	4	4	4	3	3	3
	I felt very tired	-	4	4	-	3	3	-	3	3	-	3	3	4	4	4	-	3	3
Sleep	I had problems with my sleep	4	4	4	4	4	4	3	3	3	4	4	4	3	4	3	4	3	3

Response options are frequency (F) for most of the items - none of the time, only occasionally, some of the time, often, most or all of the time; D – no difficulty, slight, some, a lot of difficulty, unable; D1 - severity S – none, mild, moderate, severe, very severe; S1 – not at all, a little bit, somewhat, quite a bit, very much.

FV – face validity, PV – psychometric validity, OV – overall validity.