EEPRU

Title: Development and Validation of the Recovering Quality of Life (ReQoL) Outcome Measures

Anju Keetharuth, John Brazier, Janice Connell, Jill Carlton, Elizabeth Taylor Buck, Thomas Ricketts and Michael Barkham

RR00050

March 2017

Correspondence to: Anju Keetharuth, ScHARR, University of Sheffield,

Email: d.keetharuth@sheffield.ac.uk

THE UNIVERSITY of York

The Policy Research Unit in Economic Evaluation of Health and Care interventions is funded by the Department of Health Policy Research Programme. It is a collaboration between researchers from the University of Sheffield and the University of York.

The Department of Health's Policy Research Unit in Economic Evaluation of Health and Care Interventions is a programme of work that started in January 2011. The unit is led by Professor John Brazier (Director, University of Sheffield) and Professor Mark Sculpher (Deputy Director, University of York) with the aim of assisting policy makers in the Department of Health to improve the allocation of resources in health and social care.

This is an independent report commissioned and funded by the Policy Research Programme in the Department of Health. The research was also part-funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH). http://www.clahrc-yh.nir.ac.uk. The views expressed are not necessarily those of the Department.

Acknowledgements

The authors would like to thank all the participants in the project, the staff who have been involved in the recruitment of participants, all the members of the governance groups. Suzanne Heywood-Everett and Laura Hancox for carrying out the interviews with South Asian and Polish service users, Katrin Conway and Catherine Acquadro from MAPI Group for the translatability assessment. We are grateful to the core members of the Scientific Group and those who have attended the various Scientific Group meetings. The core members of the Scientific Group were: Paul Blenkiron, Jed Boardman, Suzanne Heywood Everett, Andrew Grundy, Rob Hanlon, Jo Hemmingfield, Andrew Papadopoulos, Dan Robotham, Diana Rose and Mike Slade. Others who attended Scientific Group meetings were: Rosemary Barber, Laura Hobbs, Adrian Phillipson, Jenny Trite and Lisa Doughty. We also thank the members of the Psychometrics Advisory Group: Jakob Bue Bjorner, Tim Croudace and John Browne. Special thanks to Andrew Grundy, Rob Hanlon, Jo Hemmingfield, John Kay, Dan Robotham and Diana Rose, for providing us with service users' perspectives.

Ethical approval was obtained from the Edgbaston NRES Committee, West Midlands (14/WM/1062). For the online panel in the validation study, ethics approval was obtained from the School of Health and Related Research Ethics Committee via University of Sheffield Research Ethics Committee (Re: 010299).

Governance permission was obtained from each of the participating NHS Trusts. Informed consent was obtained from all participants in the study.

EXECUTIVE SUMMARY

This report presents the development and initial validation of the 10-item and 20-item self-reported, recovery-focused quality of life outcome measures named Recovering Quality of Life (ReQoL). The development of the ReQoL was commissioned by the Department of Health as an instrument to measure outcomes of mental health service users aged 16 and over in primary, secondary and tertiary care settings. The ReQoL measures are recovery focussed as they adequately capture all the key themes associated with recovery.

Methods

We adopted a mixed methods approach in a four-stage process. We employed qualitative methods for theme identification, item development (stage I) and initial testing with over 76 service users aged 16 and over (stage II). We adopted quantitative methods for establishing dimensionality through exploratory and confirmatory factor analyses, item reduction and scale construction using Item Response Theory and classical psychometrics with data obtained from >6500 service users (stage III). A selection of items for 10 and 20-item versions synthesised psychometric evidence from item response theory models and classical psychometrics, as well as qualitative evidence (stage IV). Reliability was examined by Cronbach's alpha, test re-test reliability coefficients, validity via convergence, known group differences using standardised effect sizes, and responsiveness using standardised response means.

Results

ReQoL-10 and ReQoL-20 contain positively and negatively worded items which reflect one physical health item and the following six mental health quality of life themes: activity, hope, belonging and relationships, self-perception, wellbeing, and autonomy. Unidimensionality was established showing that the items measure one construct of mental health quality of life and recovery. Both versions of

the ReQoL achieved acceptable internal consistency (alpha >.92) and test-retest reliability (>.85). They reflected known group differences (general population versus patient population and the severity of non-psychotic conditions), convergence with related measures, and were responsive over time (SRM>.4). ReQoL-10 and ReQoL-20 performed marginally better than SWEMWBS and markedly better than EQ-5D.

Conclusions

The ReQoL measures offer a number of important advantages over existing measures. They are the only ones known to the authors that have been built around the themes of recovery. The measures have been co-produced with service users who have been involved in all stages of the project as participants, advisers, partners and decision-makers. By virtue of the high face and content validity which is a result of co-production with service users as well as its brevity and good psychometric properties, ReQoL-10 has the potential to be a useful outcome measure for use in routine clinical practice. ReQoL-20 can also be used in routine practice mainly at assessment and end of treatment but also in research studies.

The measures are now available for use and licences can be obtained from the University of Oxford Innovation Ltd.

Contents

List of figures

List of tables

Acronyms

CAMHS	Child and Adolescent Mental Health Services
CFA	Confirmatory Factor Analysis

CFI	Comparative Fit Index
CHIME	Connectedness, Hope, Identity, Meaningful activity and Empowerment
CORE-10	Clinical Outcomes in Routine Evalutation-10
CPPP	Care Pathways and Packages Project
CROM	Clinician Rated Outcome Measure
DH	Department of Health
DIF	Differential Item Functioning
DSM	Diagnostic and Statistical Manual (of Mental Disorders)
EAP	Expected *a Posteriori*
EEPRU	Policy Research Unit in Economic Evaluation of Health and Care Interventions
EFA	Exploratory Factor Analysis
EUG	Expert User Group
GAD	Generalized Anxiety Disorder
GAD-7	Generalized Anxiety Disorder-7 item
GP	General Practice
IAPT	Improving Access to Psychological Therapies
ICC	Intraclass Coefficient
IRT	Item Response Theory
LOWESS	Locally Weighted Scatterplot Smoothing
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
OCD	Obsessive Compulsive Disorders
PAG	Psychometrics Advisory Group
PHQ-9	Patient Health Questionnaire-9 item
PREM	Patient Reported Experience Measure
PROM	Patient Reported Outcome Measure
QALY	Quality Adjusted Life Year
QoL	Quality of Life
ReQoL	Recovering Quality of Life
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
SES	Standardised Effect Size
SRM	Standardised Response Mean
SWEMWBS	Short Warwick and Edinburgh Mental Well-Being Scale
WLSMV	Weighted Least Squares Means and Variance Adjusted

1. Introduction

1.1 Background

In December 2013, the Department of Health (DH) commissioned the Policy Research Unit in Economic Evaluation of Health and Care Interventions (EEPRU) to construct a Patient Reported Outcome Measure (PROM) to assess quality of life in people with mental health problems. The measure was to be generic in nature and developed from the input of service users at all stages of the research. It was to cover two super clusters: psychotic and non-psychotic disorders (1). The non-psychotic clusters include common mental problems (e.g. depression, anxiety, obsessive compulsive disorders (OCD), phobias), severe and complex non-psychotic problems (e.g. personality disorder), and the psychotic clusters covering schizophrenia and other psychotic problems. The plan was to construct two versions of the PROM. The long version would be suitable for use in clinical decision-making and the shorter measure would be suitable for routine use. These measures would be scored using psychometric methods and preference weights generated to allow the computation of Quality Adjusted Life Years (QALYs) for use in economic evaluation. The measures, which are currently both known by the umbrella term Recovering Quality of Life (ReQoL), would be suitable for use in primary, secondary and tertiary settings with those aged 16 and over. To date, the ReQoL measures are mainly for self-completion but interviewer and proxy versions may become available in the future. The long and short versions of the ReQoL are expected to comprise about 20 and 10 items respectively.

1.2 Rationale for a new measure

The DH is committed to increasing the use of outcome measures in the National Health Service (NHS) to improve care for people with mental health problems. The NHS Outcomes Framework aims to incorporate indicators to promote a greater focus on the outcomes that matter most to patients. Therefore, the DH commissioned the Mental Health Foundation to examine what should be the

main domains for measuring recovery outcomes in mental health. This resulted in a report by Boardman *et al*. (2) entitled 'Assessing recovery: seeking agreement about key domains'. This report identified subjective measures of individual recovery which reflected key areas of personal recovery. The authors argued that, among other tools, a PROM was required that covered the key components known as CHIME: Connectedness, Hope, Identity, Meaning and Empowerment (3). These themes were subsequently confirmed through a review of the qualitative literature on quality of life in mental health service users and through further primary work with mental health service users (4-6). Boardman (2) stated that "For routine use more work is required to develop and test out possible (PROM) measures" (p. 3).

At the same time, NHS England had been developing a national tariff for implementation in mental health. Through this arrangement, trusts would be paid for the services rendered using the national tariffs. A key part of this initiative would require services to measure user satisfaction through a Patient Reported Experience Measure (PREM) and clinical outcomes through a Clinician Reported Outcome Measure (CROM) and a Patient Reported Outcome Measure (PROM). A review undertaken for the Quality and Outcomes work stream for the Care Packages and Pathways Project (CPPP) concluded that the more established PROMs in mental health are typically too focused on specific diagnoses or patient groups and/or are symptom focused, for example Patient Health Questionnaire-9 (PHQ-9) or Generalized Anxiety Disorder-7 (GAD-7) (7). The Short-Warwick Edinburgh Mental Well-Being Scale (SWEMWBS) was chosen for testing in secondary mental health services because its brevity and generic content made it potentially suitable across all clusters (7). The results of the pilot study were mixed with professional staff who were more sceptical about the use of the measure than service users. While it was feasible to collect the SWEMWBS once at baseline, it proved very difficult to obtain follow-up measures and there was insufficient data to analyse sensitivity to change in the various clusters (8).

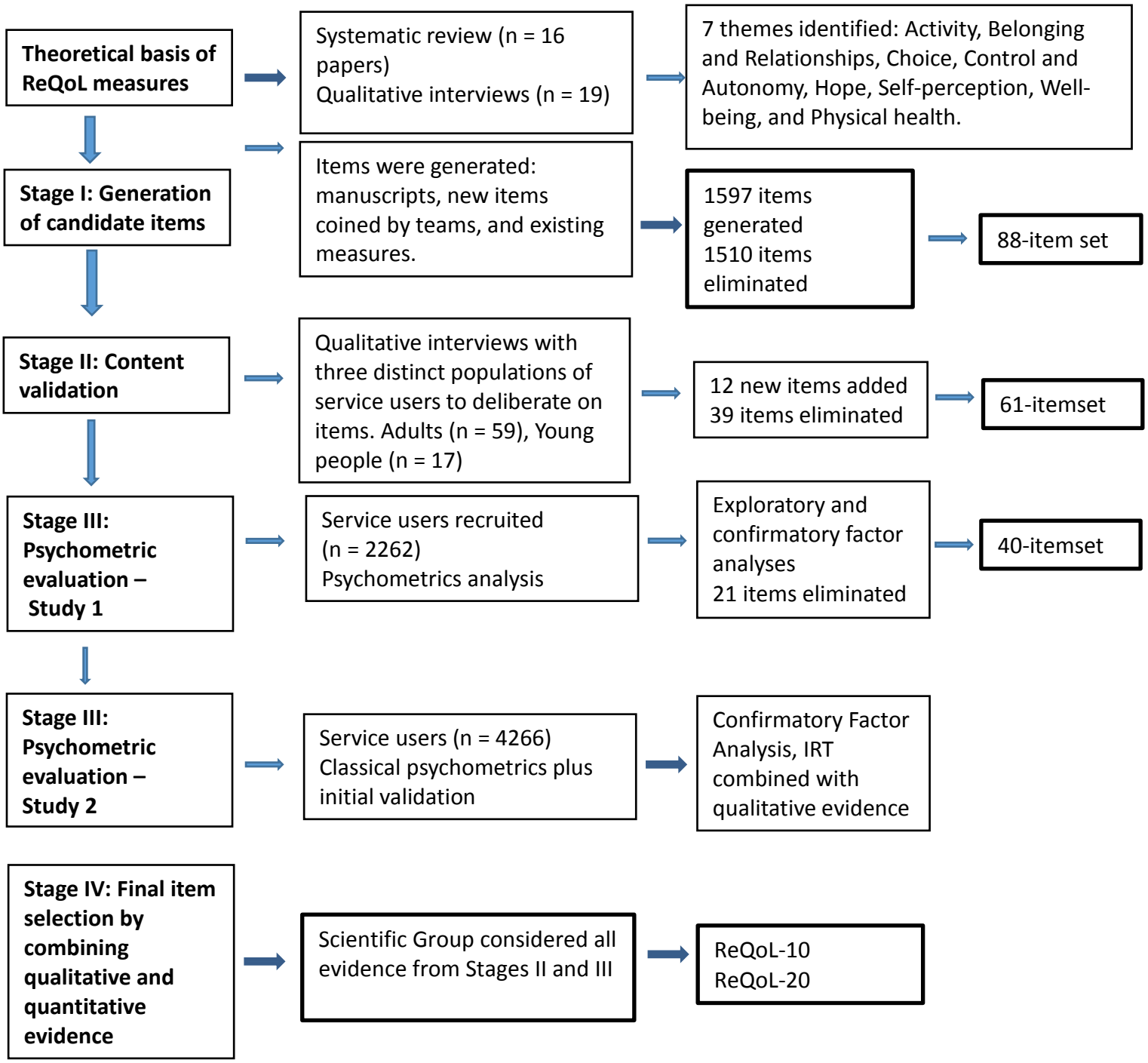
The WEMWBS (14 item version) was developed for the general population and as a result focused on the presence of positive well-being. It was inspired by some of the design principles, constructs and content from the Affectometer 2, though it included only positive items (9, 10). People with mental health problems, however, do not only identify with things that add to quality of life, but also those that take quality of life away. A systematic review of the qualitative literature and interviews with mental health service users found that the mental health domains contain both negative and positive aspects (4, 5, 11). There is also evidence that negative components, such as negative social relationships, have a greater impact on quality of life and recovery than positive ones (12). It is important for a measure to be sensitive to both improvement and deterioration in quality of life across all levels of severity. For example, it might be argued that the absence of a positive affect, such as not feeling optimistic about the future, is not the same as feeling hopeless (a negative affect). Logical, semantic and empirical arguments or explorations have not always converged, but it seems likely that reductions in hopelessness might reflect an important improvement which may not initially be associated with increased optimism and so would not be picked up by an exclusively positive item measure. Another example would be that "not feeling happy" is not necessarily the same as "feeling depressed". The absence of negative aspects may limit the relevance of the WEMWBS as a PROM of recovery to mental health service users.

From the perspective of the National Institute for Care and Health Excellence (NICE), there are limitations of using EQ-5D in calculating Quality Adjusted Life Years (QALYs) in cost-effectiveness analyses. It has been recognised that the EQ-5D, which is the main generic measure used in the DH PROMs programme, is not appropriate in many areas of mental health (4, 13, 14). Therefore, there is an appetite for a preference-based generic measure more relevant to mental health service users.

1.3 Overview of the project

There are four main stages in the development process of the ReQoL measures. They are summarised in Figure 1.

Figure 1 Development of the ReQoL



The fifth stage of the project involves the elicitation of preference weights from the general public of selected items from the ReQoL-10 measure on a scale of zero (dead) to one (full health). Once the set of weights is available in the autumn of 2017, ReQoL can be used to generate QALYs in economic evaluation of interventions in the area of mental health.

1.4 Governance groups

A key concern in the development of the ReQoL measures was that scientific rigour should be combined with the views of service users and clinicians to achieve the aims of the measure. To address this concern, four governance groups were created. The Scientific Group (15 members) was the decision-making group; the Advisory Group comprised mainly of academics (33 members) whose role was to comment on all the stages of the project and provide specific advice as and when required; the Stakeholder Group (32 members) was predominantly made up of policymakers, clinicians and NHS employees, and their role was to comment on all stages of the project; and the Expert Users Group was a group of service users and researchers with service users' perspectives who were involved in decision-making of and commenting on all aspects of the project. Expert users were also members of the three other governance groups. After data collection, the Psychometric Advisory Group (PAG) was formed to advise on methods, modelling and to discuss aspects of interpretation that would be important for selecting the different item sets.

At the end of each stage a consultation was prepared alongside a screencast presenting the results from that stage (available on http://www.reqol.org.uk/p/screencasts.html). Members of the governance groups were provided with an opportunity to comment on various issues online or by email. Separate phone calls were arranged with members if it was deemed necessary. The comments are collated and feed into the final decisions at each stage of the ReQoL project.

1.5 Criteria informing the design of the ReQoL measures

Six criteria informed the design of the ReQoL measures. The first and most important criterion being that they were based on the outcomes which service users identify as being most central to them in recovering their quality of life, rather than symptoms. The other six criteria that informed the design were that they should be: available in a version that was short enough for initial assessment and repeated use in routine outcome measurement settings but with a longer version or item set for research purposes; suitable for use with a wide spectrum of mental health conditions and levels of severity; appropriate for individuals aged 16 and over; suitable for self-completion; and free to publicly funded service delivery organisations.

1.6 Contents of this report

This report provides an account of the four stages of the development process of the ReQoL measures and the initial validation of the two ReQoL measures: ReQoL-10 and ReQoL-20. Chapter 2 provides the theoretical basis of the ReQoL measures where the seven themes are identified. Chapter 3 reports the first stage of the process relating to the shortlisting of potential items. In Chapter 4 (stage II), we report the face and content validity of the item sets following individual interviews and focus groups with adults and younger individuals aged 16-18, consultations with clinicians, interviews with a cross-cultural subset of the population, and a translatability assessment. In Chapter 5, we report two psychometric analyses undertaken in the development process (stage III). Chapter 6 relates the process of combining the qualitative and quantitative evidence to produce the final ReQoL measures. The initial validation results are presented in Chapter 7 followed by a discussion and conclusion in the final chapter.

2 Theoretical basis of ReQoL measures

The development of the themes was mainly based on work undertaken by Connell and colleagues (4-6).

2.1 Systematic review

A literature search was carried out on primary qualitative research studies (involving methods such as interviews and focus groups) which explicitly asked adults with mental health problems what they considered to be important to their Quality of Life (QoL) or how their QoL had been affected by their mental health problems. It identified 7,000 studies from which 200 full texts were retrieved. A total of 13 studies were relevant for this study and they were all set in developed countries: Canada, UK, Sweden, USA, Australia and New Zealand. There was a slight bias for occupational therapy and nursing. The main limitation was that the papers focused on severe mental health conditions, especially schizophrenia. A synthesis of the data was carried out using a "framework" approach for the analysis of primary data. It is a highly structured approach to organising and analysing data which permits the expansion and refinement of an a priori framework to incorporate new themes emerging from the data.

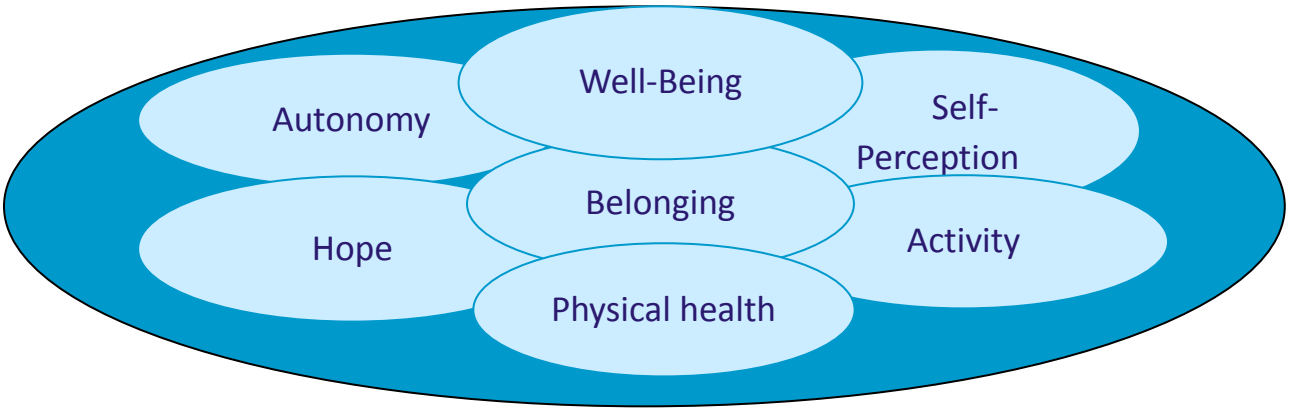
2.2 Qualitative interviews

The aim of the interviews was to broaden the range of diagnoses and severity covered in the review. Nineteen interviews were conducted with service users in both primary and secondary care. In the secondary care setting, people with severe mental health problems (for example schizophrenia, personality disorder) were recruited from two community mental health teams. Those with mild to moderate problems were recruited from Improving Access to Psychological Therapies services. The major limitation was the absence of any service users with OCD or bipolar disorder. Among those with schizophrenia, all were male. Framework analysis was used to allow the identification of common and variable patterns of themes.

2.3 Themes

The review identified six themes and the interviews fitted this very well except for physical health which was an additional concern arising from the interviews (Figure 2). These were adopted for ReQoL.

Figure 2 Themes that matter to the quality of life of mental health service users



As shown in Figure 2, there is some overlap within the themes and they are therefore not mutually exclusive.

2.4 Conceptual overlap

The overlap between QoL and Recovery arose quite early as among the papers retrieved there were more recovery-based papers than QoL ones. For pragmatic reasons, the recovery papers were excluded from the study. However, there is significant overlap with the themes that Leamy *et al.* (3) identified; they map onto each other, as shown in Table 1, and overlap with other concepts such as subjective experience and subjective well-being which also looked at psycho-social needs.

Table 1 Quality of Life versus recovery framework

Quality of Life (Connell *et al.*)	Recovery (Leamy *et al.*)
Belonging/Relationships	Connectedness/support/stigma and discrimination/community participation
Hope/Hopelessness	Hope
Self-perception	Identity
Activity (meaningful/enjoyable)	Meaning
Autonomy/Choice/Control	Empowerment
Well-Being/Ill-Being	Well-Being/Symptoms
Physical Health	

As shown in Table 1, apart from the physical health domain, the mental health domains identify very similar themes that matter to people with mental health conditions even though they have been termed slightly differently. There is some overlap within the domains and they are therefore not mutually exclusive.

The conceptual overlaps encountered include:

1. An overlap between QoL and Recovery. This issue arose quite early as among the papers retrieved there were more recovery-based papers than QoL ones. For pragmatic reasons, the recovery papers were excluded from the study. We will return to this point later.
2. There is overlap with other concepts such as subjective experience and subjective well-being which also looked at psycho-social needs. There is also some overlap with the work being done by the Mental Health Foundation's Strategies for Living.
3. Another issue faced is that QoL and Recovery are seen as positive concepts.

2.5 Positive and negative sub-themes

In the recovery framework domains are seen as positive concepts. However, in the systematic review and the interviews (4-6) it was clear that there were a lot of negative concepts, for example service users talk in terms of the absence of depression or anxiety. Therefore, the domains can be thought of in terms of both enhancing Quality of Life and reducing Quality of life. An example here might be stigma which takes away from people's Quality of Life (Table 2).

Table 2 Themes and sub-themes

	Themes	Positive sub-themes	Negative sub-themes
1.	Well-being	Happiness	Depression/sadness
		Relaxed/Calm	Fear/anxiety/worry
		Positive energy/motivation	Loss of energy/motivation
2.	Sense of Belonging/ Relationships	Sense of belonging	Not belonging - outsider
		Positive relationships	Negative relationships
		Friendship and camaraderie	Lack of friends
		Supportive relationships	Unsupportive relationships
3.	Activity/Meaning	Enjoyable activity	Unenjoyable/ stressful activity
		Meaningful/valued/purposeful /constructive	Boring/meaningless/not valued
4.	Self-perception/Identity	Positive self-identity	Negative self-identity
		Positive self-confidence	Negative self-confidence
5.	Autonomy/Control/ Choice/Empowerment	Autonomy	Dependence
		Choice	Lack of choice
		Control/coping	No control/not coping
6.	Hope	Hope	Hopelessness
		Plans and goals	Lack of plans and goals
7.	Physical Health	Good health	Poor Health

3 Item generation (Stage I)

Once the domains for the ReQoL measure had been agreed with the Scientific Group, the first stage was to generate a set of potential items to best represent each of the themes and sub-themes. This section describes the sources used to create an item pool and the process that has been adhered to in an attempt to retain the most appropriate items.

3.1 Sources for generating pool of potential items

To generate a pool of items, three main sources were used: 21 commonly used quality of life, psychiatric risk, distress and severity measures were consulted with a total of 523 items shortlisted; 17 existing recovery measures with a total of 580 items extracted; and the transcripts from Connell *et al*. (6) identified a total of 494 items capturing the language of service users.

3.2 Criteria used in shortlisting items for the ReQoL measures

The process of selecting items from a total of 1,597 was to use a set of criteria adapted from those originally proposed by Streiner and Norman (15) (Table 3) and a number added by the research team as shown in Table 4

Table 3 Criteria adapted from Streiner and Norman (15)

1.	Reading Level	Rule of thumb: reading skills should not exceed those of a 9 year old
2.	Ambiguity	Poorly worded items. Even straightforward items may pose a problem if not applicable, e.g. *I like my spouse* is problematic if someone does not have a spouse.
3.	Double-barrelled question	This is 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*.
4.	Jargon	The vocabulary should not be technical and should be part of everyday vocabulary.
5.	Value-laden words	Judgmental statements may prejudice the respondent and should therefore be avoided (e.g. *having more social contact may not be seen to be better by everyone*).

6.	Positive and negative wording	Negatively worded items should be avoided, e.g. it is better to have the item “I feel ill most of the time” than “I rarely feel well”.
7.	Length of items	The items should be as short as possible but not so short that it loses comprehensibility.

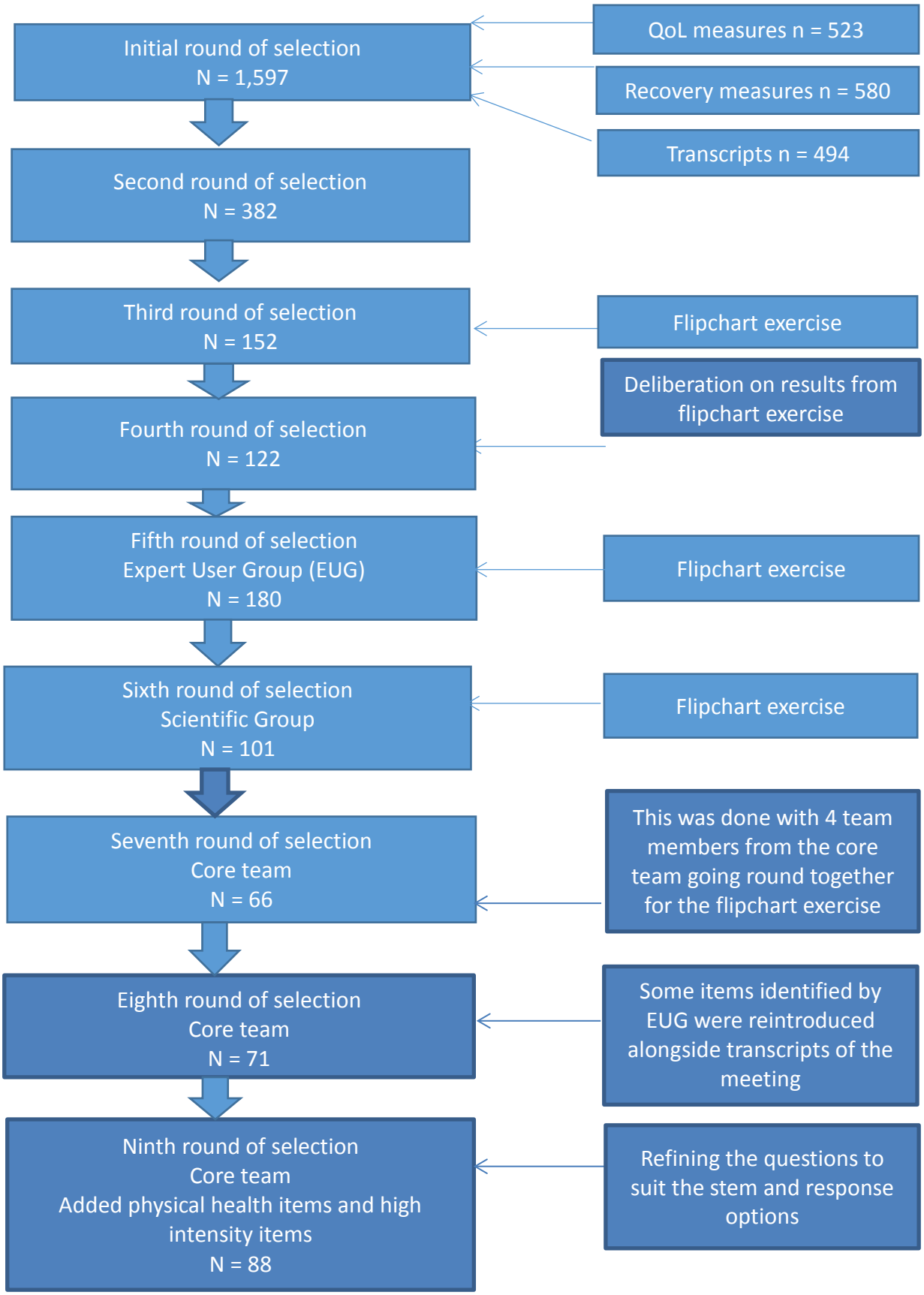
Table 4 Further criteria added by the ReQoL Research Team

8.	Too specific to a lifestyle	E.g. referring to employment when some respondents may not work.
9.	Too specific to a diagnosis	E.g. hearing voices is too specific.
10.	Makes comparisons over time	E.g. “compared to usual” is not appropriate for long term mental health problems
11.	Does not lend itself to short time periods	E.g. things that are more intrinsic personality traits and very unlikely to change.

The initial identification of the item pool was done by two researchers, Jill Carlton and Janice Connell. The item reduction process is summarised in Figure 3. In a series of core team meetings, each theme was reviewed and each item was discussed and retained if it met the criteria outlined in Tables 4 and 5. This process brought the number of items from 1,597 to 382. At that point, all the 382 items were transcribed on sticky notes, displayed on a flipchart and stuck on the walls of the room. Each team member walked round and allocated “votes” to the most appropriate items by inserting a sticker next to their preferred items. The votes for each item were counted and used to inform the item selection.

Similarly, this exercise was undertaken by the Expert User Group and the Scientific Group. The votes were counted and further deliberation in round seven reduced the number of items to 66. In the eighth and ninth rounds physical health items were added, additional items were reintroduced from the Expert User Group (EUG) suggestions, and some high intensity items were reintroduced to ensure that the items covered the breadth of severity of mental health conditions. At the end of Stage 2, there were 87 items that were shortlisted and retained for use in Stage 2 to test for content validity.

Figure 3 Stage 2 – item reduction flowchart



4 Qualitative evidence: content validation of the ReQoL (Stage II)

4.1 Introduction

Content validity is defined as the extent to which the set of items comprehensively covers the different components of health to be measured (16). Face validity, often viewed as a part of content validity, considers whether the items of each domain are sensible, appropriate and relevant not only to the researchers developing the measure, but also to the people who use the measure on a day-to-day basis (17). It is important, therefore, that these tests are carried out during the development of the measure and not only after the measure has been finalised. In the past, the content and face validity of many outcome measures were based on the judgements of researchers and health care professionals, with little or no input from service users (18). However, it is now recognised that these judgements should be informed by service users (19). There is evidence to suggest that what may be regarded as a good outcome by a clinician or researcher may differ from what is important to a service user (20). Ensuring that a measure reflects the outcomes that are important to service users requires active input from service users in the development stages of the measure (21).

In this chapter we report on the second of four stages in the development of the ReQoL. The aim of this stage was to assess the acceptability of the items and their validity to service users who represent those likely to be completing the questionnaire.

4.2 Methods

A qualitative study using face-to-face structured interviews and focus groups with service users was undertaken.

4.2.1. Recruitment

There was a requirement that ReQoL be suitable for mental health service users over the age of 16. Therefore adults (aged 19-79) and young adults (aged 16-18 years) from NHS mental health services and a local charity were invited to participate. In order to obtain views from across the spectrum of mental health service users, broad inclusion criteria were applied; the only exclusion concerned people experiencing acute episodes of their mental health condition, those not well enough to take part, and those who could not speak English or give informed consent. This allowed for maximum variation of mental health problems, the severity of problems, and current service contact.

Adult participants were recruited from four UK NHS Trusts providing mental health services and a UK charity in the North of the country. Two Trusts were located in the South of England and two in the North. Recruitment was undertaken by health care staff and clinical studies officers within the individual Trusts. The recruitment of young adults (aged 16-18) took place from two further NHS Trusts based in the Midlands and the North of England and was undertaken by the interviewer (Child and Adolescent Mental Health Service clinician), health care staff, and a clinical studies officer. A pragmatic sampling approach was adopted which differed according to the usual practices of the participating Trusts. Some Trusts held lists of volunteers who had given their permission to be approached. Others recruited participants when seen at individual or group therapy sessions, approached potential volunteers on the wards or recruited from an active service user group. Most Trusts used multiple methods.

4.2.2 Participants

A total of 59 adult service users took part; 40 participated in individual interviews, 11 attended two focus groups of seven and four participants respectively, and four interviews took place where two participants were present at the same time, as these service users requested that they not be interviewed alone. Interviews lasted between 20 mins and 1 hr 40 mins for the adult sample.

A total of 17 young adults participated; 15 participated in individual interviews and two participants chose to be interviewed together. Interviews lasted between 30 mins and 1hr 13 mins for the young adult sample.

Table 5 Characteristics of participants in Stage II

		Adults		Young Adults	
		N	%	N	%
Gender	Male	22	37.3	4	23.5
	Female	37	62.7	12	70.6
	Transgender	0	0	1	5.9
Age	16-18	0	0	17	100
	19-29	12	20.7	0	0
	30-39	15	24.1	0	0
	40-49	12	20.7	0	0
	50-59	10	17.2	0	0
	60-69	6	10.3	0	0
	70+	1	1.7	0	0
	Not indicated	3	6.8	0	0
Ethnicity	White British	46	77.6	15	88.2
	Black/Black British	6	10.3	0	0
	Asian/Asian British	2	3.4	0	0
	Mixed/Multiple ethnic group	2	3.4	2	11.8
	Other ethnic group	3	5.2	0	0
Employment/Activity	Employed	16	25.9	2	11.8
	Student	2	3.4	15	88.2
	Retired	4	6.9	0	0
	House-person	2	3.4	0	0
	Not in employment	25	43.1	0	0
	Other	8	13.8	0	0
	Not indicated	2	3.4	0	0
Diagnosis/Own View*	Schizophrenia/Psychosis	26	44.8	0	0
	Depression	27	46.6	6	32.3
	Bipolar	8	13.8	0	0
	Anxiety Disorders	21	24.4	4	23.5
	Eating Disorder	2	3.4	2	11.8
	Personality Disorder	4	6.9	0	0
	Other	4	6.8	0	0
	Not indicated	2	3.4	5	29.4
Current Care**	None	3	5.2	0	0
	General Practitioner	6	10.3	1	5.9
	Improving Access to Psychological Therapy	8	13.8	0	0
	Community Mental Health Team	29	50.0	0	0
	Child and Adolescent Mental health Services (CAMHS)			14	82.3
	CAMHS Inpatient			2	11.8
	Adult Inpatient	8	13.8	0	0
	Voluntary sector	3	3.4	0	0
	Not indicated	2	3.4	0	0

* Some participants provided more than one diagnosis/opinion of their presenting problem

** Some participants indicated that they were receiving current care from more than one service provider

4.2.3 Interviews

Interviews and focus groups were conducted between September 2014 and June 2015. All interviews and focus groups were undertaken at NHS sites familiar to the participants. Participants were provided with an information sheet prior to the interviews. Written consent was obtained at the time of the interview prior to any data collection. Participants were asked to complete a short demographics form indicating their gender, age, ethnicity, employment, education level, mental health diagnosis and their own perception of their mental health problem (which may or may not be the same as the diagnosis). Interviews and focus groups with adult service users were conducted by three experienced qualitative researchers, one of whom is a service user (Janice Connell, Jill Carlton, Andrew Grundy). Interviews with young adults were undertaken by an experienced qualitative researcher and clinician who specialises in child and adolescent mental health (Elizabeth Taylor Buck). All participants were given a £20 shopping voucher in recognition of their time.

The interview process was the same for both adult and young adult service users. To help establish the validity of each potential item, participants were asked whether it was meaningful and relevant to their quality of life; whether it was clear, understandable and easy to answer; and the reason they either liked or disliked an item, or preferred it to another. They were also asked for their preferred item within a group or pair of items thought by them to relate to a similar concept, and where it was likely that only one item would be included in the reduced item pool relating to that concept (e.g. between *"I felt relaxed"* and *"I felt calm"*). Alternative wordings to items were elicited if participants thought an item was important to their quality of life but unclear or difficult to answer.

An iterative approach was undertaken with the interviews. Adult participants were initially presented with a set of 88 potential items separated into the themes previously identified by service users (via a systematic review of qualitative literature and service user interviews) as being important to quality of life (*well-being*, *relationships and a sense of belonging*, *activity*, *self-*

perception, *autonomy*, *hope*, and *physical health*). At approximately the half way stage, after interviews had been undertaken at two of the Trusts, 12 items (which were primarily a reformulation of existing questions) were added as a result of the feedback provided by those participants who had been interviewed. This increased the item set to 100 potential items for the remaining interviews. As a result of the findings from the adult participants (which are reported in this paper), a meeting was held with scientific and user group members and decisions were made to remove some items from the set and to change others. This reduced the number of items to 61. Young adults were presented with this reduced item set, which meant there was less potential response fatigue for these participants. Due to the large size of the item set, not all participants provided their views on all items. The items were presented in a different order to reduce any fatigue effect, but kept within their respective themes. The comfort and enthusiasm of the interviewee to continue with the interview was considered at all times. The majority of the interviews were recorded and these recordings were transcribed. Notes were taken for the three adult interviewees who preferred their interview not be recorded, and were also included in the analysis. All identifying information was removed from the transcripts and notes prior to analysis.

4.2.4 Analysis

From the transcripts, the comments made by each participant for each item were charted into a spreadsheet framework with items across the horizontal axis and participants on the vertical axis. A traffic light system was used to highlight negative (red), positive (green) and neutral or ambiguous comments (orange). A thematic analysis of the comments was undertaken to establish the underlying reasons for the popularity, or lack of popularity, of the items. This information was used as a starting point for discussion with the scientific, advisory and service user expert groups to establish whether or not an item should remain as a potential item in the ReQoL measure. In the final stage of the development of the measure (Chapter 6), this information was used in conjunction

with other evidence, such as the psychometrics and feedback from clinicians, to make a decision on the final items.

In the results section, the quotes are labelled as follows: A-Adult, YP-Young Person, ID number, I-Individual, and F-Focus Group.

4.3 Results

The following five themes were identified as contributing to the rejection or acceptance of an item: *relevance of items*, *ease of response*, *item ambiguity*, *potentially distressing items*, and *judgemental items.*

4.3.1 Relevance of items

There were some items which were universally liked and had few objections. The commonality between these items was that the service users could relate to each of them as being something they commonly experienced. Therefore, the items were relevant to their own mental health and, in turn, their quality of life. As a result, the responses to these questions did not require much thought and were considered easy to answer by the vast majority of participants. Items which fell into this category were "*I had difficulty getting to sleep or staying asleep"*, "*My health limited day to day activities"*, "*I felt able to trust others*", "*I felt anxious*" and *"I felt confident in myself"*. These items were considered to be particularly relevant because of the impact these experiences or feelings had on other aspects of their quality of life.

(I felt able to trust others)

"A really good question - if you're feeling not so good or a little bit paranoid or whatever then you know the trust definitely goes down. And I just think that's important as well

because I think if you don't feel like you can trust other people then you're certainly not going to feel very happy." (AI24)

(I felt confident in myself)

"You need confidence to be able to value yourself - when you have got no confidence and you are down you don't value nothing." (AF3)

Some items were described by the service users as being irrelevant or meaningless, either to their own mental health problems or to quality of life. For example, service users felt that is was more important to quality of life that they accepted themselves rather than be accepted by others. It was felt that it was not necessary to "*feel loved*" for a good quality of life, and was perhaps a bit of a luxury, but feeling *"cared for"* was important and less specific to a certain type of relationship. It was also felt impractical and unachievable to have everything under control or be able to do all the things you wanted to.

(I felt loved)

"I think to feel loved might be a luxury - but to feel cared for erm might be sufficient. I mean I don't suppose the health service can erm prescribe or give love but they can provide care." (AI6)

(I could do all the things I wanted to do)

"I don't think the next one is very valid - I can't do all the things I want to do - swim with dolphins, it just isn't going to happen (laughs). No." (AI12)

There were objections to some items because they were thought to be too general; they could be applied to anybody and therefore were not specific to mental health difficulties. Examples included:

"*I avoided things I needed to do*", it was felt that there may be very good reasons to avoid doing things you did not want to do, and doing so could enhance your mental health; "*I felt irritated*" which was felt to be a normal reaction and not necessarily linked to mental health; and "*I felt tired and worn out*" which could be due to physical health as well as mental health.

> *(I felt tired and worn out)*
>
> *"Feeling tired and worn out can be like fatigue or lots of different things […] a lot of teenagers feel like tired and worn out all the time because it's just kind of how we are, sort of thing, but like if like everything is an effort, like even like brushing your hair is an effort, that's sort of like different to feeling a bit tired." (YPI14)*

A few of the young adults thought that being "confused about who you are" was a natural thing for individuals in their age group.

> *"I think that you're going to get that with a lot of people really and I think it just like develops, like as you get older you tend to get less-, […] because at that time you've got lots of like you've got a lot of hormone changes and imbalances in your body and it's just you get mood swings all the time and it's just like you're going through lots of different things and it can just be…" (YPI3)*

4.3.2 Ease of response

Some items were considered difficult to answer by some participants because they were too abstract, thus requiring too much thought. Whilst under normal circumstances this would be fine, it was something they felt could be particularly problematic while in a distressed state when first accessing mental health services. This particularly applied to items where they would be required to consider the thoughts of others, such as: "*I thought people did not understand me*", "*I felt*

discriminated against", "*I felt accepted as who I am*" and "*I thought people did not want to know me*". Rather than it being about whether they thought people understood them, some would answer literally and try to think about whether people did understand them. Some participants disliked this added layer of complexity.

> *(I thought people did not want to know me)*
>
> *"I don't know what other people think. 'Didn't want to know me', what do you mean by that? … It's very vague isn't it, sort of verging on a paranoid thought isn't it." (AI34)*

> *(I felt discriminated against)*
>
> *"I mean I would think if I was answering this and I would look back I would think 'I don't know', I think it's a hard thing to say that someone has discriminated against you." (AF5)*

Other items were thought by some service users to be vague or abstract, and therefore difficult to answer, such as: "*I felt trapped*", "*I felt at ease with who I am*", "*I had a purpose in life*", "*I neglected myself*", "*I thought I could make positive changes*", "*everything in my life felt bad*" and "*I felt confused about who I am.*"

> *(I felt confused about who I am)*
>
> *"It's like sort of, what you are like confused about? Yourself, as in like your sexuality or things like that, or confused about yourself as in like 'am I really who people seem to think I am', or even like down to like thinking that you are someone else, entirely like a different person like people are lying to you about who you are and stuff, so I think it depends a lot on maybe your condition or even like your own personality on how that would be interpreted […] I think it is a kind of confusing question." (YPI14)*

There were some items that the service users thought may be difficult to provide an honest answer to, because of their own low level of self-awareness whilst ill. It was sometimes only in retrospect that they may realise they had "*neglected themselves*" or were not "*thinking clearly*".

(I neglected myself)

"I think that when I have been neglecting myself I wouldn't have known so like maybe I hadn't showered for like three weeks but I probably wouldn't realise that I was neglecting myself." (I31)

(I was thinking clearly)

"At the time you think you are thinking clearly, especially if you are in an episode of psychosis what is in your head is very real so I think people could probably answer yes to that one." (AI26)

4.3.3 Item ambiguity

Some items could be interpreted in more than one way, for example whether an item related to physical or emotional health. "*I was in pain*" was considered by some to be about emotional rather than physical pain, and there was some ambiguity over whether "*I had problems with self-care, washing or dressing*" was related to emotional or physical problems, or both.

(I had problems with self-care, washing or dressing)

"Good one - because it is not necessarily physical - it can reflect a person's mental capacity. I've been asked before have you got any problems with self-care, washing, dressing, and I go no, and they say well you know you once didn't get dressed for two months, you lived in those pyjamas remember. And I think oh yes, but because it would be under a physical

heading I would associate it with I've got problems with self-care due to physical disability." (AF5)

There were items that were intended to be negative (i.e. indicative of a poor quality of life) but in some circumstances this could be regarded as positive. For the item *"I felt guilty"*, a couple of the interviewees indicated that this could be a positive change when they had done something regrettable whilst ill.

(I felt guilty)

" […] *there are two types, there is a good one and bad one, it is good to have guilt because it shows that you are a decent person… it is a good thing, it is another thing that shows you are getting better cos when I was committing my crimes, I didn't think that I had done owt wrong because I weren't well like, and when I got well the guilt kicked in, so I think a guilty sign is when you are getting well - but to have too much it can ruin your life can't it?" (AF3)*

Similarly, for the item "*I did things I found stressful"* it was pointed out that there are some people who thrive on stress, thus for some it could be positive. There were also items which were intended to be positive but could be interpreted negatively. For example "*I had a purpose in life"'* when the intention was to end their life, "*thinking clearly"* can be due to a total lack of emotion, and doing something "bad" could be enjoyable or rewarding.

(I did things I found rewarding)

"I think… erm... they could have punched someone in the face do you know what I mean. They could have done something that they think is big and hard or they think that it's the right thing to do but it's not." (YPI2)

For the most part, participants felt that having choices was a good thing, however some acknowledged that having too much choice could be detrimental.

> *(I had choices about what I did)*
>
> *"I mean the sort of the school of erm mental health often says the reason people get illnesses is because there's so much potential in your life to do things… I mean we get TV showing us celebrities and role models and when people see these things they think they can do the same but they can't… and that's why people get disillusioned… so I think you can have too many choices." (AF6)*

For young adults, being in control was not necessarily indicative of good quality of life, and the item about "coping" was preferred.

> *"Coping is getting through it, just getting through it, but controlling is being in charge […] the control may be too much for some people. Do you know what I mean?" (YPI2)*

The negative interpretation of intended positive items could occur with participants with bipolar disorder. Feeling "happy", "full of life" or "safe" could be due to the experience of a manic episode.

> *"* […] *because I have got bipolar so if I was in a manic phase I feel safe… I don't see that there is anything dangerous… I could be in danger and not know I am in danger."*
>
> *(Interviewer: "Relating that to Quality of Life, is that a good thing or a bad thing?")*
>
> *"It's a bad thing, but I feel like it's a good thing." (AF5)*

Other items which had problems with interpretation were: "I felt discriminated against", which could be thought of in association with race or gender rather than mental health; "I had reasons to get out

of bed in the morning", which did not necessarily mean they were happy or that they did actually get out of bed; and for the item "I felt hopeless" it was felt that "hopeless" had the two interpretations "lack of hope" and "useless".

4.3.4 Potentially distressing items

One of the common reasons given for objecting to an item, and having the view that it should not be included in a quality of life measure, was that it would cause upset. Some items were considered to be *too* negative. These were often related to suicidal thought and intent. It was the extreme negative (and direct) wording within items that participants found distressing, such as *"I had thoughts of killing myself"*, and "*I thought I would be better off dead"*. The wording was described as being *"upsetting"*, "*harsh"* and "*too strong",* and could provoke suicidal thoughts. Conversely, other participants did not object to a direct approach as they were used to answering such questions. The majority of participants thought that suicidal intent was an important indicator of quality of life and that it was also important to identify it to their key mental health worker. Those participants who objected to the direct terminology expressed a preference for items with a more indirect, sensitive approach. The items *"I did not care about my own life"* and *"I thought my life was not worth living"* were preferred.

> *(I had thoughts about killing myself)*
>
> *"I don't like 'killing myself', I don't like that expression, I will be honest with you, maybe 'have you had thoughts about self-harm', could it be worded another way? I would agree it is important to ask because lots and lots of people do have suicidal thoughts." (AI15)*

> "*When I was younger it might have upset me to think about suicide [but]… once you have been through it all, the idea of people killing themselves is not really upsetting at all." (YP)*

Other items considered to be too extreme by some participants, and thus described as upsetting, related to feelings about the self. The items *"I felt humiliated or shamed by other people"*, *"I felt useless", "I felt shame", "I felt stupid"* and *"I detested myself"* were described as "too personal", "embarrassing", "not nice" and "traumatising". One participant stated that such a negative question about the self would make "the voices" more prominent. For items relating to the self, the positive items were preferred (for example *"I felt confident in myself" or "I felt ok about myself"*). Of the negative items a gentler approach was favoured, for example *"I disliked myself"* was much preferred to "*I detested myself"*.

> *(*I detested myself*)*
>
> *"I think I would be knocked back by it I think, I would be like 'oh do I', you know and then it's not, it brings on feelings like 'well yes I am this and I am that'… I think it might be too strong a word because it might bring back awkward feelings." (AI26)*

> *"'Detest' - it might be a bit embarrassing to say and dislike is a bit of an easier one to swallow." (AI4)*

Due to its sensitive nature, a number of people responded that they would not like to admit to certain things and therefore would either not respond to the item, or would not answer honestly. The reasons given were that they would find it embarrassing *("I felt humiliated or shamed by other people"*, *"I had problems with self-care, washing or dressing"*), they had concerns surrounding the consequences of disclosure (*"I had thoughts about killing myself"*), or both (*"I have threatened or intimidated another person"*).

> (I had thoughts about killing myself)

"*If you went to a nurse here and said that you had thoughts about killing yourself then you would be on numbers where you would be-, just have somebody watching you constantly, they watch you sleeping, so I think personally people would not answer that." (AI2)*

*(*I have threatened or intimidated another person*)*

"That's quite extreme isn't it so I wouldn't like to answer that question." (AI38)

(I had problems with self-care, washing or dressing)

"I wouldn't like to answer that, it's making me feel a bit ashamed that I can't take care of myself." (AI36)

Some items were felt to be insensitive because they were *too* positive, in particular the items "*I felt full of life",* "*I felt I could bounce back from my problems*", and to a lesser extent "*I felt happy*". Participants felt these to be unrealistic, as they thought they were never likely to feel this way even when they were well. Such items were described as "*patronising*" and "*daft*", particularly if asked when they were accessing a mental health service for the first time when they would be feeling particularly distressed. When further explanation was given that the questionnaire would measure change from the beginning to end of therapy, the latter was felt to be more appropriate as they may have short periods of happiness.

*(*I felt full of life*)*

"I think it's that thing where people might think 'well I wouldn't be here if I was full of life and happy'... I think it's got to have an air of realism about it. For the person who's reading it to think that it's a good reflection of how they're feeling or you know you've understood their situation. I just think that they're a bit too wahooo, here's your party banner and your balloon kind of thing." (AI24).

4.3.5 Judgemental items

Some positively worded items were thought to be too judgemental and reflected an opposing value system. This particularly applied to those items that were related to doing "good" things: "*I was able to do things that helped others*", "*I felt I made a contribution"*, "*I did things that I found worthwhile"* and "*I felt useful"*. It was felt by some participants that helping others was not necessary for quality of life and that people with mental health problems were not necessarily in a position to be able to help others or to do things that were worthwhile, and doing such activities could make them feel worse. Furthermore, participants noted that if the items were answered truthfully (that they did not do things that helped others) this may result in feelings of guilt.

> *(I felt I made a contribution/I was able to do things that helped others)*
>
> *"Is that a sign of being well, you could just be a selfish person? For some people it wouldn't be part of their life to do something to help other people so there is no relevance. My charity work helps me and helps others - but that is more to do with my work ethic than wanting to help others. If I saw something like that in a questionnaire when I was feeling very poorly it would just make me feel worse because to me that looks like a judgement. You could be making a contribution but it's actually making you more ill so is it a positive thing or is it actually a negative thing?" (AF1)*

Conversely, many participants stated that doing positive things did make one feel better, and were important to quality of life.

> *(I was able to do things that helped others)*
>
> *"That's very important and it makes you feel nice." (AI4)*

The concept of "independence" was also thought to be judgemental because of the assumption that independence was something to be valued. The addition of "*as I would like to*" aimed to address this issue. The rewording was liked by some participants but ignored by others. Some considered the additional wording made the question too complex and difficult to answer.

> *(I lived as independently as I would like to)*
>
> *"I'm not sure. I'm not sure if living independently is an imitation of erm good mental health… and I think there's an implication with being independent that you're doing alright and if you're not, you're not." (AI6)*

Participants liked items that they felt were non-judgemental, and items that they would have no objections to admitting to or answering honestly.

4.4 Convenience sample to assess response options

To assess the response options for the ReQoL questionnaire, interviews were carried out with a convenience sample of non-academic staff at the University of Sheffield. Fourteen interviews were carried out using the severity scale (n = 8) and frequency scale (n = 6) to test how well the response options worked with the items shortlisted. The main problems with the severity scale were:

- It was difficult to answer the questions about the last week.
- The labels "moderately" and "extremely" were often seen as unrealistic or vague.
- Several items did not make sense using the severity responses (e.g. There were people I could turn to for help).

On the whole, the frequency scale was more acceptable to participants. As a result, it was decided that a frequency scale would be used for the mental health items, though the severity scale would be retained for the physical health items. Agreement responses (yes, no or degree of agreement) were disregarded as an option due to the difficulties they would pose in the valuation stage of the project.

4.5 Consultation with clinicians – parallel study

As part of the consultation process, feedback on the 40- item set (used in Study 2 in Chapter 5) was obtained from a group of 11 clinicians working for two mental health service providers. Additionally, focus groups were carried out with 35 clinicians including staff from all the main professional groups involved in multidisciplinary mental health care from six different providers. The focus groups addressed the relevance, clarity and usefulness of each item from a clinician perspective. Clinicians were asked to select a preferred question or questions from each theme, and provide reasons why these questions were preferred. The staff focus groups identified two or more preferred items in each theme, together with detailed pros and cons regarding each of the 39 mental health items. There was a large degree of agreement between focus groups regarding problematic questions, which informed final item selection. For instance, in the item *Everything in my life felt bad*, the word "everything" made it difficult to answer the question as it could be that one person is having only one thing that felt bad but its impact could be substantial.

4.6 Cross cultural issues with Urdu and Polish speaking service users – parallel study

Two focus groups and two individual interviews were carried out with Urdu speaking service users in Bradford. One focus group and one individual interview were also carried out with Polish speaking service users. The questions were translated into Urdu and Polish but English versions were also available at the interviews. Participants were asked to reflect on the relevance and wording of 61 items. There were a few emerging issues from interviews with South Asian service users such as the importance of family, stigma, social and family judgement and how one is perceived by others. The Polish speaking service users came to England shortly after the Second World War and they found it difficult to relate to the terms like "worry" as they perceive the focus of their lives to be more about coping and survival.

4.7 Translatability Assessment

Finally, a translatability assessment following established guidelines (22) was carried out to identify potential semantic and structural issues that might be a barrier to future translations of items in the measures.

The qualitative results from all the service users and the translatability assessment were combined to further reduce the number of items to 61. The results from the adolescents and young adults, the cross cultural studies, and the clinicians were used to inform the final selection of the ReQoL items as discussed in Chapter 6.

5 Psychometric evaluation (Stage III)

This chapter presents the findings from the psychometrics analyses from two quantitative studies (Study 1 and Study 2) where service users were asked to complete the ReQoL item sets.

5.1 Methods

The aims of Study 1 were: a) to explore the dimensionality of the item set, and b) to identify items that could potentially be excluded because of redundancy in order to lessen the response burden. The aims of Study 2 were: a) to replicate the dimensionality results of Study 1, and b) to perform a more in-depth evaluation of item performance to inform the final item selection for the measures.

5.1.1 Participants

There were 2,262 and 4,266 service users in Studies 1 and 2 respectively. Participants comprising both inpatients and outpatients were recruited from 13 and 20 secondary mental health providers in Study 1 and Study 2 respectively. There were also participants from three General Practices and a trial cohort in each study, and three and two voluntary sector organisations in Studies 1 and 2

respectively. In Study 1, a total of 500 participants were recruited from an online panel. To maximise response rates, a combination of modes of recruitment was used. Participants were recruited face-to-face while attending services, some completed the survey by post, and others online. Table 6 presents the demographics of participants from both studies. In Study 1, participants completed the 61-item set at one time point only. In Study 2, participants completed a reduced 40-item set (see below), of whom 953 completed a follow-up 6 to 12 weeks later. Participants in Study 2 also completed a second measure in order to assess convergent validity, the results of which are reported in Chapter 6.

Table 6 Characteristics of the samples recruited in the psychometric testing stages

	Stage IV – Study 1 (n = 2,262) Mean (SD) or n (%)	Stage IV – Study 2 (n = 4,266) Mean (SD) or n (%)
Age categories		
16 to 25	261 (12)	441 (10)
26 to 64	1,541 (68)	2,879 (67)
65 and over	390 (17)	681 (16)
Missing	687 (3)	265 (6)
Ethnicity		
White	1,932 (85)	3,649 (86)
Non-white	296 (13)	384 (9)
Missing	32 (1)	233 (5)
Diagnoses		
Common mental health disorders	794 (35)	1,423 (33)
Schizophrenia	213 (9)	421 (10)
Other psychotic disorders	116 (5)	234 (5)
Bipolar	201 (9)	411 (10)
Personality Disorder	106 (5)	238 (6)
Others	239 (11)	252 (6)
Missing	593 (26)	1,287 (30)
Setting in which recruitment occurred		
GP practices	145	1,146 (27)
Improving Access to Psychological Therapies (IAPT)	-*	261 (6)
Secondary care - outpatients	1,285	1976 (46)
Secondary care - inpatients	64	563 (13)
Community	565	310 (7)
Life satisfaction - How satisfied are you with your life nowadays? Score 0 - not satisfied at all to 10 - extremely satisfied.	5.4 (2.7)	5.3 (2.8)

**In Stage 4, participants from IAPT had been recruited but classified as secondary care outpatients.*

5.1.2 Statistical methods

5.1.2.1 Factor analyses (Studies 1 and 2)

Confirmatory Factor Analysis (CFA) was employed to understand whether the mental health themes identified in Stage I represented distinct underlying constructs and their inter-relationships. Items concerning a physical health theme were excluded from the factor analyses as physical health was deemed a priori and conceptually to be a different construct. Second, Exploratory Factor Analysis (EFA) using Geomin rotation was performed to identify other potential factors present in the data. Finally, informed by the EFA, single factor, two-factor and bifactor, CFA models were estimated. As the item responses were captured on a 5-point Likert scale, the variables were treated as ordinal categorical. CFA was performed using the robust weighted least squares means and variance adjusted (WLSMV) estimator (23). The factor analyses were carried out in Mplus 7.4 (24). Model fit was assessed by the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI). For the RMSEA and CFI, a value of ≤0.8 and >0.95 was assumed to provide a good fit respectively.

5.1.2.2 Item Response Theory (IRT) analyses (Study 2)

Graded response models (GRM) (25) were used for all analyses. Model fit was evaluated by the sum-score based item fit statistic (S-G^2) (26). Since the S-G^2 statistic is calculated for each item, the approach may lead to spurious results in cases of large numbers of items. To counter this problem, Study 2 was divided into 4 datasets (N >1,000 each). A sample size of a minimum 1,000 was considered sufficient to identify relevant misfit. Only items with misfit ($p<0.05$) in three to four datasets were considered misfitting. After fitting the IRT models, item and test information functions were examined. Information functions indicate the precision of measurement for people at different levels of severity on the latent scale and are dependent on the item parameters. All IRT analyses used IRTPRO 3.0 (27).

Differential item function (DIF) with regards to age, gender, ethnicity, and diagnosis was evaluated through ordinal logistic regression models (28). Significant DIF was assessed through a dual criterion of statistical significance and a difference in explained variance (Nagelkerke pseudo R^2) larger than 2% (29).

5.1.2.3 Sensitivity to change

Sensitivity to change was examined at item level in terms of the correlation between the change scores for each item and the global assessment of change at follow-up. It was calculated using the Wilcoxon matched pair signed rank test (30) in STATA 14 (31).

5.2 Results

Psychometric evaluation - Study 1

In the initial CFA, the six mental health factors did not provide a satisfactory model and the factors were strongly correlated. The results from the EFA of the mental health items suggested a two-factor solution (eigenvalues for the first 5 factors in Study 1 were 29.2, 3.1, 1.6, 1.2, and 1.1). All the negatively worded items (n = 34) loaded on the first factor and all the positively worded items (n = 23) loaded on the second factor. The correlation between the two factors was 0.8. The CFA results are presented in Table 7. Redundancy found in the factor analysis results in Study 1 were combined with the qualitative evidence on the items from Stage II in order to reduce the item set from 61 to 40 items (see Figure 3). This 40-item set comprising 39 mental health items and one physical item was retained for Study 2.

Table 7 Fit statistics from confirmatory factor analytic models

	6 factor model		2 factor model: negative, positive		Bifactor model: global, negative, positive	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
Chi Square Value	21,576	26,483	13,093	13,317	12,859	11,224
DF	1,524	687	1,538	694	1,482	647
RMSEA (<0.08)	0.091	0.095	0.069	0.066	0.070	0.062
CFI (>0.95)	0.921	0.937	0.955	0.969	0.955	0.974
WRMR	3.486	6.015	2.475	2.978	2.260	2.304

The items deleted from the item set following the results of Study 1 were:

Table 8 Items deleted using Study 1 data

Theme	**Item**
Activity	My mental health limited day to day activities
	I found things interesting
Belonging and Relationships	There were people I could turn to for help
	I was able to help others
	I thought nobody cared about me
	I enjoyed being with other people
Choice, Control and Autonomy	I had choice about what I did
	I felt trapped
	I was able to cope with everyday life
Hope	I believed I could make positive changes
	My life seemed pointless
Self-perception	I felt unsure of myself
	I tended to blame myself for bad things that have happened
	I felt confused about who I am
	I felt ok about myself
Wellbeing	I had difficulty controlling my worry
	I felt tired and worn out
	I felt scared
	I had feelings of despair
	I felt miserable (was added to the set by the scientific group)

Psychometric evaluation - Study 2

For the 40-item set, missing data did not exceed 5% for any of the items as per Table 9. All the response options were endorsed and no obvious ceiling or floor effects were observed, as shown in Table 10.

Table 9 Missing data of the 40-item set by item

Theme	Item	Total sample n = 4266	
Activity	I found it difficult to get started with everyday tasks	139	3%
	I did things I found rewarding	166	4%
	I neglected myself	188	4%
	I avoided things I needed to do	173	4%
	I enjoyed what I did	164	4%
Belonging and Relationships	People around me caused me distress	145	3%
	I felt lonely	161	4%
	I felt able to trust others	171	4%
	I felt people did not want to be around me	166	4%
	I thought people cared about me	169	4%
Choice, Control and Autonomy	I could do the things I wanted to do	161	4%
	I felt overwhelmed by my problems	166	4%
	I had the opportunity to do the things I wanted	151	4%
	I felt unable to cope	175	4%
	I felt in control of my life	180	4%
Hope	I felt hopeful about my future	181	4%
	I felt hopeless	162	4%
	Everything in my life felt bad	162	4%
	I thought my life was not worth living	152	4%
Self-perception	I felt like a failure	157	4%
	I felt confident in myself	163	4%
	I felt at ease with who I am	162	4%
	I valued myself as a person	140	3%
	I disliked myself	182	4%
Wellbeing	I felt calm	137	3%
	I felt miserable	143	3%
	I felt safe	150	4%
	I was disturbed by unwanted thoughts and feelings	155	4%
	I felt irritated	165	4%
	I felt angry	154	4%
	I felt relaxed	177	4%
	I felt terrified	179	4%
	I felt everything was an effort	166	4%
	I felt panic	159	4%
	I felt happy	174	4%
	I found it hard to concentrate	159	4%
	I worried too much	161	4%
	I felt anxious	181	4%
	I had problems with my sleep	149	3%
Physical health		299	7%§

§This was higher than the rest due to the presentation of the question in the survey booklet.

Table 10 Endorsement frequency (Study 2: n = 4,266)

Item description	Levels				
	1	**2**	**3**	**4**	**5**
I found it difficult to get started with everyday tasks	589	853	1051	889	745
	14%	*21%*	*25%*	*22%*	*18%*
I felt able to trust others	466	822	991	893	923
	11%	*20%*	*24%*	*22%*	*23%*
I felt unable to cope	481	655	850	825	1,280
	12%	*16%*	*21%*	*20%*	*31%*
I could do the things I wanted to do	410	988	1,168	703	836
	10%	*24%*	*28%*	*17%*	*20%*
I felt happy	583	1,020	1,110	751	628
	14%	*25%*	*27%*	*18%*	*15%*
I thought my life was not worth living	381	446	573	610	2,104
	9%	*11%*	*14%*	*15%*	*51%*
I enjoyed what I did	452	834	1,234	752	830
	11%	*20%*	*30%*	*18%*	*20%*
I felt hopeful about my future	713	948	1,029	668	727
	17%	*23%*	*25%*	*16%*	*18%*
I felt lonely	623	699	807	777	1,199
	15%	*17%*	*20%*	*19%*	*29%*
I felt confident in myself	826	974	982	617	704
	20%	*24%*	*24%*	*15%*	*17%*
I did things I found rewarding	576	962	1,198	781	583
	14%	*23%*	*29%*	*19%*	*14%*
I avoided things I needed to do	566	810	984	834	899
	14%	*20%*	*24%*	*20%*	*22%*
I felt irritated	483	895	1,080	983	660
	12%	*22%*	*26%*	*24%*	*16%*
I felt like a failure	686	649	717	709	1,348
	17%	*16%*	*17%*	*17%*	*33%*
I felt in control of my life	803	957	903	642	781
	20%	*23%*	*22%*	*16%*	*19%*
I felt terrified	241	377	630	655	2,171
	6%	*9%*	*15%*	*16%*	*53%*
I felt anxious	868	914	824	801	678
	21%	*22%*	*20%*	*20%*	*17%*
I had problems with my sleep	1,080	766	715	716	840
	26%	*19%*	*17%*	*17%*	*20%*
I felt calm	381	964	1,256	792	736
	9%	*23%*	*30%*	*19%*	*18%*
I found it hard to concentrate	778	877	965	842	645
	19%	*21%*	*24%*	*21%*	*16%*

All further analyses used the 39 mental health items. The results from the EFA of the mental health items suggested a two-factor solution (eigenvalues for the first five factors were 24.6, 2.3, 1.1, 0.8, 0.8). All the negatively worded items (n = 24) loaded on the first factor and all the positively worded items (n = 15) loaded on the second factor with a Geomin correlation between the two factors of 0.8. A two-factor CFA model provided an acceptable fit but a bifactor model comprising a global factor and two local factors of negative and positive affects yielded a slightly superior fit in both Study 1 and 2 (see Table 7). The factor loadings on the negative and positive factors were considerably smaller than the loadings on the global factor, thereby supporting an essentially unidimensional model. Thus, in Study 2 the global factor explained 83% of the common variance, the negative factor explained 13%, and the positive factor 4%. However, residual correlations had to be modelled for 16 pairs of items (ranging from 0.24 to 0.47 in a standardised solution).

In the IRT analyses conducted in Study 2, two items were found misfitting in three of the datasets: *I felt at ease with myself* and *I could do the things I wanted to do*. The marginal reliability for response pattern scores of the 39 items was 0.98.

Items were sensitive to change in the right direction. In other words, participants who reported an improvement and a worsening in overall health also reported higher and lower quality of life in their ReQoL scores respectively. Two items did not show a statistically significant change (5%) in the much better category: *I thought people cared about me* and *I felt angry*.

5.3 Scoring of the 40-item set

Two methods of scoring were compared. From the IRT analyses, expected *a posteriori* (EAP) scores were calculated that estimate the expected value of the probability distribution of latent trait scores (32, 33). Baseline IRT scoring was compared with summative scores for the whole sample and also

for primary versus secondary care participants. Changes in scores computed from IRT scoring were also compared with changes from summative scoring. While IRT scores and sum scores were strongly correlated (r = 0.98) both for baseline/initial assessment and repeated/follow-up ReQoL, there were noticeable differences for some participants. The correlation between change in sum scores and change in EAP scores was 0.95. IRT scoring did not provide any benefit in terms of providing more statistical power when scores for primary and secondary care service users were compared. Correlations between the scores were calculated alongside bubble plots to illustrate the differences (Figures 5-8).

Figure 4 Comparing IRT scoring with summative scoring at baseline

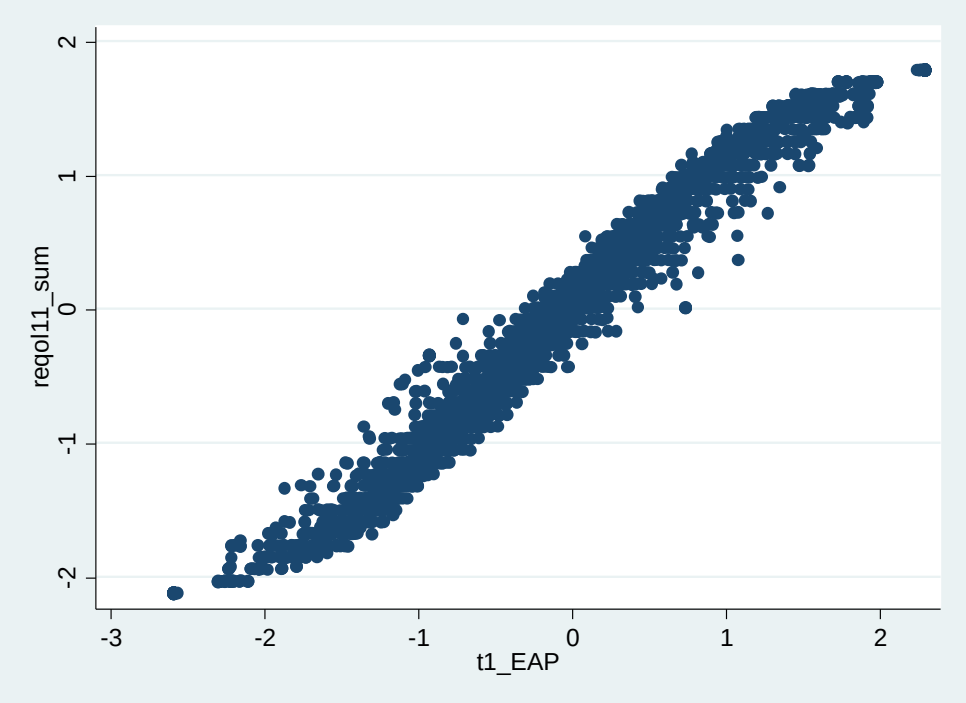
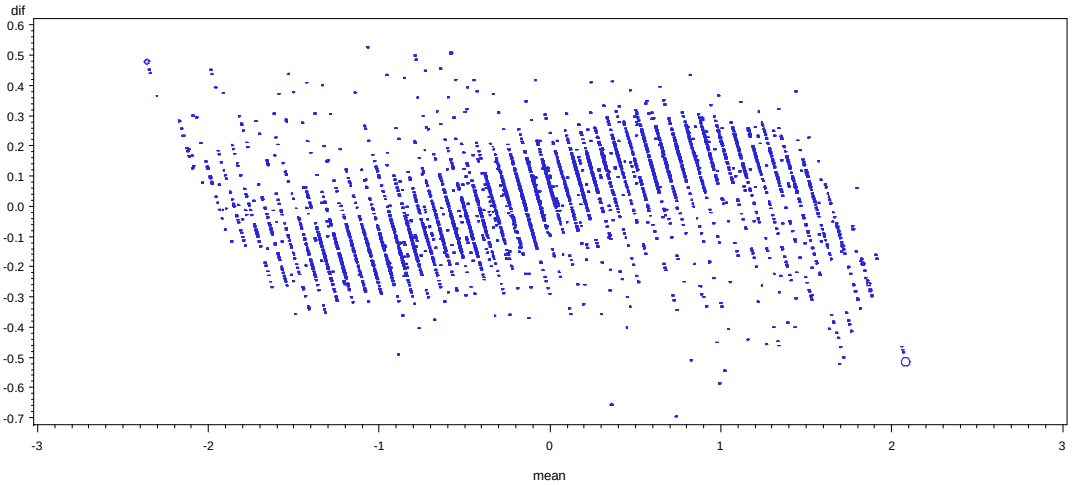


Figure 5 Bland Altman bubble plot of the difference (sum score minus EAP score) vs the mean.



The difference (sum score minus EAP score) can be up to slightly over 0.5 SD.

At follow-up, participants were asked a global question to assess the change in their health status. The plots in Figures 7 and 8 are classified by those who reported that their health improved a lot or somewhat (Improved), deteriorated a lot or somewhat (Worsened), or remained the same (About the same). The difference is <0.5 SD.

Figure 6 Comparing changes in EAP and sum scores by those whose health has improved, worsened or stayed the same.

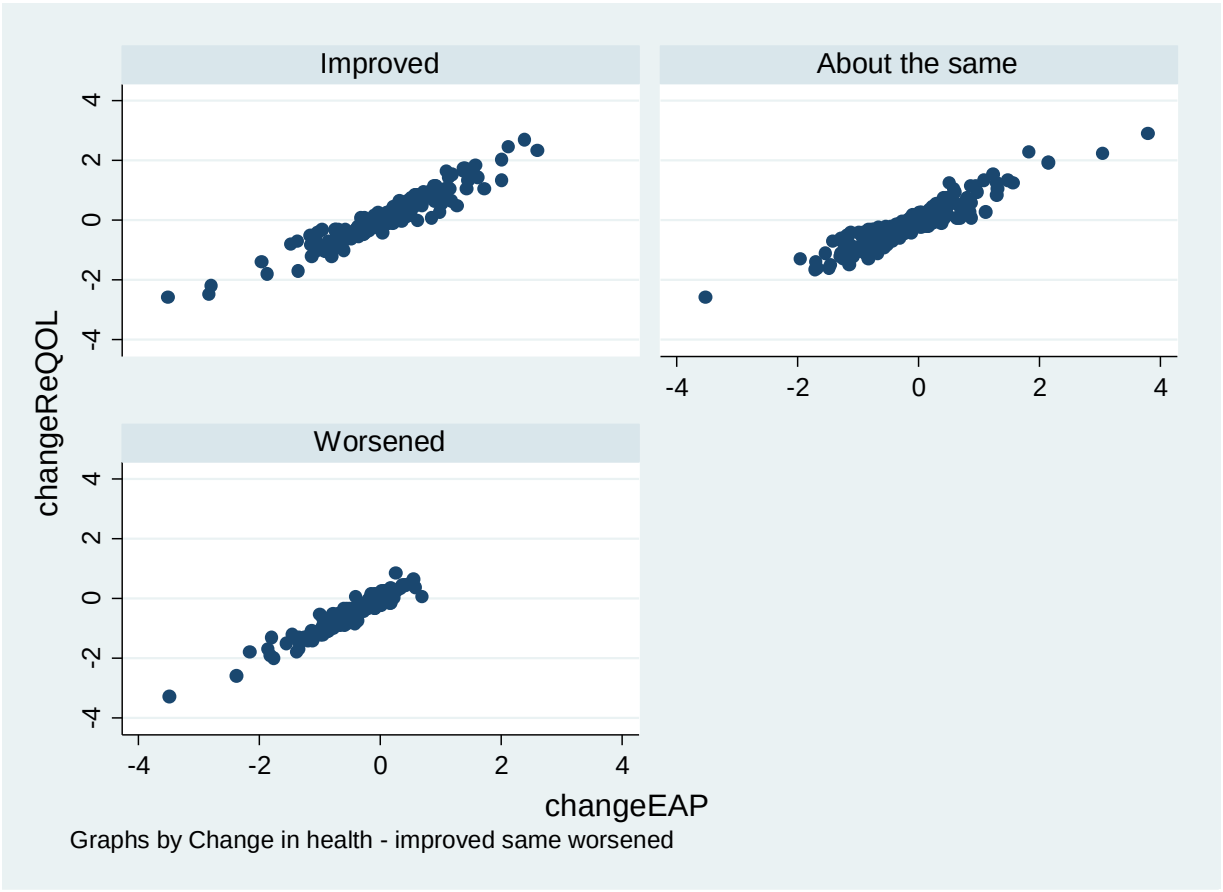
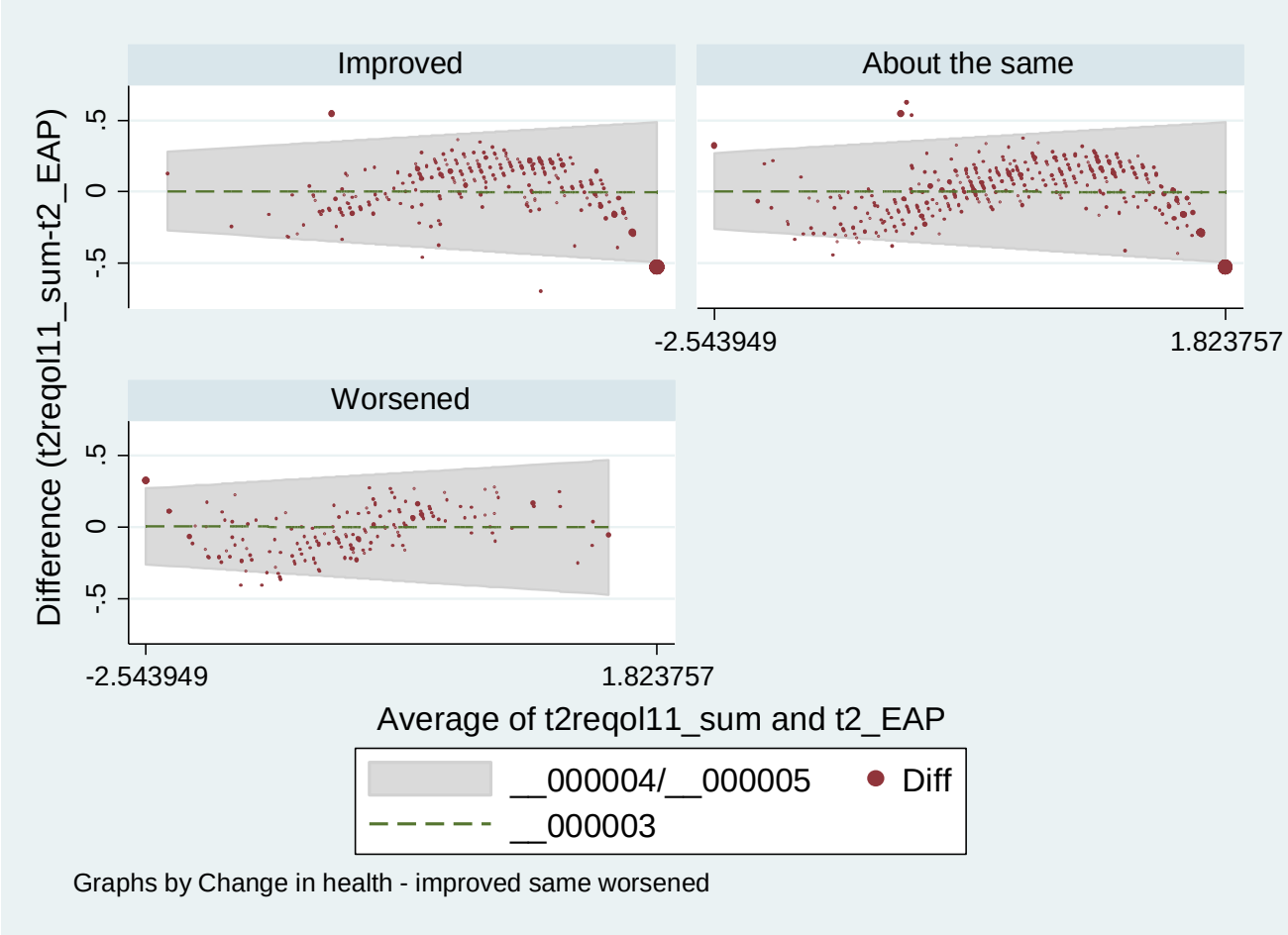


Figure 7 Bland Altman plot of the difference (change sum score minus change EAP score) vs. the mean



In theory, the IRT score is the best available estimate of the true score. However, it comes with a price of complexity in the scoring procedure. As a result, the summative score is recommended.

6 Final item selection- combining qualitative and quantitative evidence (Stage IV)

This chapter explains how qualitative evidence (Chapter 4) and quantitative evidence (Chapter 5) have been combined in a novel fashion to assist in the final selection of items for ReQoL-10 and ReQoL-20.

6.1 Guiding principles for item selection

For each item, the quantitative and qualitative evidence was synthesised to ensure that the best and most acceptable items for the two versions of ReQoL were chosen. To assist with the selection of items, the following guiding principles were identified:

At the domain/thematic level:

1) All six mental health themes should be represented in the final two versions of ReQoL to ensure that the resulting measures are recovery-focused.

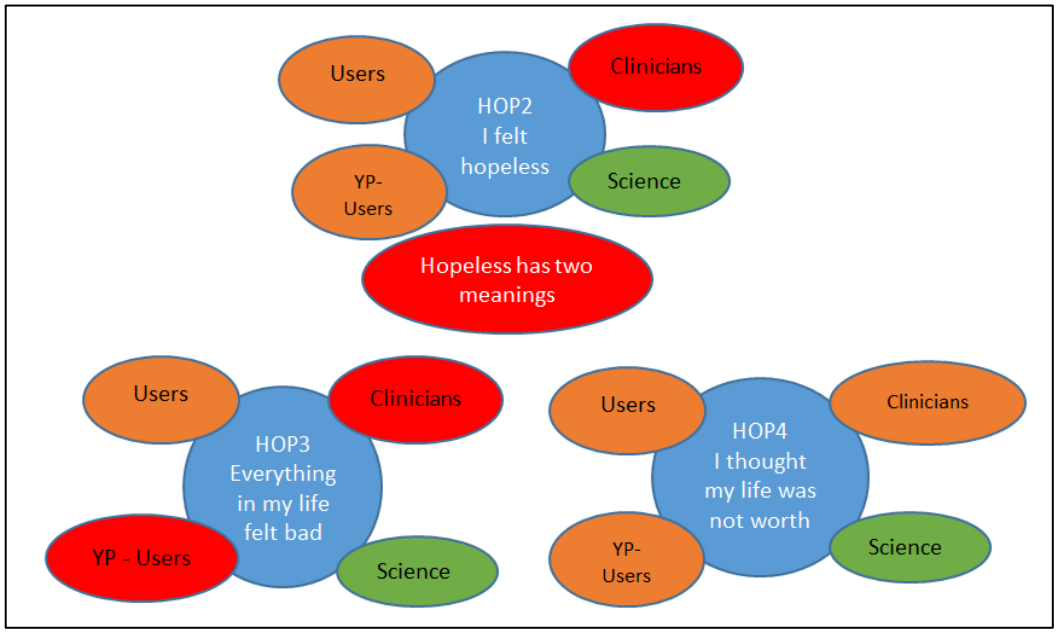
At the item level, items should:

2) Be acceptable to service users based on data collected through interviews and focus groups and during the online consultation processes.
3) Have clinical usefulness to practitioners based on data collected through focus groups with clinicians.
4) Demonstrate performance according to the following criteria in conjunction with the PAG:
    a) Absence of floor or ceiling effects,
    b) No redundancy as indicated in CFA models by high residual correlations with already selected items,
    c) No evidence of misfit in IRT-based item analyses,
    d) Contribution to measurement precision across a broad range of mental health as assessed by item information functions,
    e) No DIF,
    f) Sensitivity to change.

6.2 Final item selection process

Final item selection was undertaken in a collaborative exercise at a Scientific Group meeting. The combination of the qualitative and quantitative evidence is summarised in Table 11. To assist members in their decisions and deliberations, the qualitative (service user and clinician) and quantitative sources of evidence (IRT, sensitivity to change) were colour-coded and presented in a simple diagram (Figure 8). The process started with choosing the best item from each health domain. Once one item per domain was selected, a decision was taken as to whether a second item was needed for that theme and if so, the same process was repeated to decide on the second item.

Figure 8 Combining qualitative with quantitative evidence (theme: hope subtheme: hopelessness)



Key: Users – adult service users; YP–Users – service users aged 16-18; Science – psychometrics results
Green – evidence in favour; Amber – mixed evidence; Red – evidence against

6.3 Combining qualitative and quantitative evidence

The results from Stages I to III were used in this final stage. For illustrative purposes, Table 11 provides a summary of the findings for the 40-item set. Applying these results, the following decisions were made to construct ReQoL-10 and ReQol-20.

Table 11 Summarising qualitative and quantitative evidence for each ReQoL item by theme

Domain	Variable name	Item	Positive (Pos)/ Negative (Neg)	Accepted by users (U) / clinicians (C) /Young People (YP)	Local correlation with	IRT misfit	Centre of information	Range with item information above 0.5	Sensitive to change	ReQoL form
Activity	act1	I found it difficult to get started with everyday tasks	Neg	U/C/YP	act4, wb9	-	-0.10	-2.2 to 2.0	+	10/20
	act2p	I did things I found rewarding	Pos	U/C/YP	act5p	-	0.00	-2.3 to 2.3	+	20
	act3	I neglected myself	Neg	U/YP	act4	-	-0.71	-2.6 to 1.1	+	
	act4	I avoided things I needed to do	Neg	U/C/YP	act3, act1	-	-0.19	-2.2 to 1.9	+	20
	act5p	I enjoyed what I did	Pos	U/C/YP	act2p	-	-0.22	-2.4 to 1.9	+	10/20
Belonging	bel1	People around me caused me distress	Neg	U/C/YP	bel3p, bel5p, wb5, wb6	-	-0.56	-2.7 to 1.4	+	
	bel2	I felt lonely	Neg	U/C/YP	-	-	-0.30	-2.2 to 1.5	+	10/20
	bel3p	I felt able to trust others	Pos	U/C/YP	bel1, bel4, bel5p	-	-0.31	-2.4 to 1.7	+	10/20
	bel4	I felt people did not want to be around me	Neg	U/C/YP	bel3p, bel5p	-	-0.60	-2.4 to 1.2	+	
	bel5p	I thought people cared about me	Pos	U/C/YP	bel1, bel3p, bel4	-	-0.55	-2.6 to 1.4	-	
Autonomy	cho1p	I could do the things I wanted to do	Pos	U/C/YP	cho3p	+	-0.28	-2.5 to 1.9	+	10/20
	cho2	I felt overwhelmed by my problems	Neg	U/C/YP	cho4	-	-0.18	-1.9 to 1.6	+	
	cho3p	I had the opportunity to do the things I wanted	Pos	U/C/YP	cho1p	+	-0.18	-2.4 to 2.0	+	
	cho4	I felt unable to cope	Neg	U/C/YP	cho2	-	-0.40	-2.2 to 1.3	+	10/20
	cho5p	I felt in control of my life	Pos	U/C/YP	-	-	0.02	-1.9 to 1.9	+	20
Hope	hop1p	I felt hopeful about my future	Pos	U/C/YP	-	-	0.00	-2.1to 2.1	+	10/20
	hop2	I felt hopeless	Neg	U/YP	hop3 ho4	-	-0.44	-2.1 to 1.2	+	
	hop3	Everything in my life felt bad	Neg	U	hop2 hop4	-	-0.55	-2.2 to 1.1	+	
Domain	**Variable**	**Item**	**Positive**	**Accepted**	**Local correlation**	**IRT**	**Centre of**	**Range with**	**Sensitive**	**ReQoL**

	name		(Pos)/ Negative (Neg)	by users (U) / clinicians (C) /Young People (YP)	with	mis fit	information	item information above 0.5	to change	form
	hop4	I thought my life was not worth living	Neg	U/C/YP	hop2 hop3	-	-0.79	-2.4 to 0.8	+	10/20
Self-Perception	sel1	I felt like a failure	Neg	U/C/YP	-	-	-0.31	-2.0 to 1.3	+	20
	sel2p	I felt confident in myself	Pos	U/C/YP	-	-	0.08	-1.9 to 2.0	+	10/20
	sel3p	I felt at ease with who I am	Pos	U/C/YP	sel4p	-	-0.06	-1.9 to 1.8	+	
	sel4p	I valued myself as a person	Pos	U/C/YP	sel3p, sel5p	-	-0.06	-2.0 to 1.8	+	
	sel5	I disliked myself	Neg	U/C/YP	sel4p	-	-0.33	-2.0 to 1.3	+	
Wellbeing	wb1p	I felt calm	Pos	U/C/YP	wb7p	-	-0.23	-2.5 to 2.1	+	20
	wb2	I felt miserable	Neg	U/C/YP	-	-	-0.22	-2.3 to 1.8	+	
	wb3p	I felt safe	Pos	U/C/YP	wb8	-	-0.69	-2.7 to 1.2	+	
	wb4	I was disturbed by unwanted thoughts and feelings	Neg	U/C/YP	-	-	-0.23	-2.2 to 1.7	+	
	wb5	I felt irritated	Neg	U/YP	wb6, bel1	-	-0.11	-2.4 to 2.2	+	20
	wb6	I felt angry	Neg	U/YP	wb5, bel1	-	-0.46	-2.6 to 1.6	-	
	wb7p	I felt relaxed	Pos	U/YP	wb1p	-	0.05	-2.1 to 2.2	+	
	wb8	I felt terrified	Neg	U/C/YP	wb3p, wb10	-	-0.99	-2.8 to 0.8	+	20
	wb9	I felt everything was an effort	Neg	U/C/YP	act1	-	-0.03	-2.0 to 2.0	+	
	wb10	I felt panic	Neg	U/YP	wb8, wb14	-	-0.64	-2.5 to 1.2	+	
	wb11p	I felt happy	Pos	U/C/YP	-	-	0.00	-2.1 to 2.1	+	10/20
	wb12	I found it hard to concentrate	Neg	U/C/YP	-	-	0.07	-2.0 to 2.1	+	20
	wb13	I worried too much	Neg	U/C/YP	wb14	-	0.20	-1.7 to 2.2	+	
	wb14	I felt anxious	Neg	U/C/YP	wb13, wb10	-	0.10	-1.9 to 2.1	+	20
	wb15	I had problems with my sleep	Neg	U/C/YP	-	-	0.14	-1.3 to 1.6	+	20

The first items chosen for each theme were: *I felt confident in myself* (self-perception), *I felt lonely* (belonging and relationship), *I felt unable to cope* (choice, control and autonomy), *I felt hopeful about my future* (hope), *I found it difficult to get started with everyday tasks* (activity theme), and *I felt happy* (wellbeing theme). The scientific group considered a second item was important for the latter three themes and, in addition to the existing criteria, these were selected to complement the first items in terms of direction, range of the item information, and compatibility with the other items. The selection of the additional ten items to constitute the 20-item version followed a similar process. ReQoL-20 items were chosen to provide more item information on important sub-themes (for example sleep, concentration and control of life). This makes little difference to the overall psychometric performance (Chapter 7).

6.4 Presenting ReQoL-10 and ReQol-20 items

ReQoL-10 contains 6 positive items and 4 negative items. ReQoL-20 is made up of the ReQoL-10 items plus 10 additional items: a total of 9 positive items and 11 negative items. Both versions contain the last physical item. Although physical health is important to the quality of life of mental health service users, it is not included in the total because it is distinct from mental health.

		Response options
1.	I found it difficult to get started with everyday tasks *	5 frequency-based options ranging from: none of the time, only occasionally, sometimes, often, most or all of the time.
2.	I felt able to trust others *	
3.	I felt unable to cope *	
4.	I could do the things I wanted to do *	
5.	I felt happy *	
6.	I thought my life was not worth living *	
7.	I enjoyed what I did *	
8.	I felt hopeful about my future *	
9.	I felt lonely *	
10.	I felt confident in myself *	
11.	I did things I found rewarding	
12.	I avoided things I needed to do	
13.	I felt irritated	
14.	I felt like a failure	
15.	I felt in control of my life	
16.	I felt terrified	
17.	I felt anxious	
18.	I had problems with my sleep	
19.	I felt calm	
20.	I found it hard to concentrate	
Physical health question	Please describe your physical health (problems with pain, mobility, difficulties caring for yourself or feeling physically unwell) over the last week.	5 severity-based options ranging from: no problems, slight problems, moderate problems, severe problems and very severe problems.

Further details, sample copies, scoring guides and requests for permissions to use the ReQoL measures are available from: The Clinical Outcomes team at Oxford University Innovation Ltd at: http://innovation.ox.ac.uk/outcome-measures/recovering-quality-life-reqol-questionnaire/

6.5 Scoring guide for ReQoL-10 and ReQoL-20

6.6.1 ReQol-10 index score

Each question is scored from “None of the time” to “Most or all of the time”. The scores are found as a subscript under each response option in the actual measure. The positively worded questions (Q2, Q4, Q5, Q7, Q8 and Q10) are scored from 0 to 4. The scores are reversed for the negatively worded items negatively worded questions (Q1, Q3, Q6 and Q9) are scored from 4 to 0. The ReQoL-10 index score can be calculated by summing the numbers for the first 10 questions. If a single question is unanswered in the first 10 questions, the mean value of the other responses can be used to fill the gap. The minimum score for ReQoL-10 is 0 and the maximum is 40, where 0 indicates poorest quality of life and 40 indicates highest quality of life.

6.6.2 ReQoL-10: handling missing data

The ReQoL-10 score can be calculated if only one of the first 10 questions is unanswered. In the case of one missing item, the mean value of the other responses can be used to fill the gap. If more than one item is unanswered in the first 10 questions, then an overall score for ReQoL-10 cannot be calculated. If respondents give two answers to a single question, we recommend that the lower quality of life response is adopted.

6.6.3 ReQol-20 index score

The ReQoL-10 index score can be calculated by summing the numbers for the first 10 questions of the ReQoL-20. If a single question is unanswered in the first 10 questions, the mean value of the other responses can be used to fill the gap. The minimum score for ReQoL-10 is 0 and the maximum is 40, where 0 indicates poorest quality of life and 40 indicates highest quality of life.

To calculate the ReQoL-20 index, the numbers for questions 1-20 can be summed. In the case of ReQoL-20, the minimum score is 0 and the maximum is 80, where 0 indicates poorest quality of life and 80 indicates highest quality of life. If two questions are unanswered between questions 1-20, the mean value of the other responses can be used to fill the gap. If more than two questions are unanswered, then the overall index score cannot be calculated.

ReQoL-10 scores generated from the ReQoL-20 measure are directly comparable to scores obtained from the ReQoL-10 version only if there is a maximum of one missing item in the first 10 items.

Work regarding the best method of handling missing data for both ReQoL versions is ongoing.

7 Validation of the ReQoL measures

This chapter reports the psychometric performance of the ReQoL-10 and ReQoL-20 items in terms of reliability, construct validity and responsiveness in mental health service users experiencing a variety of mental health conditions. Comparisons are made against commonly used outcome measures favoured by providers of mental health services and also by mental health policy makers.

7.1 Methods

7.1.1 Samples

Sample (Reliability): For the reliability assessment, a sample comprising both patients and members of the general population was recruited from an online panel through a market research company. A total of 2,000 members of the general public and 800 patients were recruited (Table 12). The general population sample recruited was representative of the UK general population based on age, gender (46% male, 54% female), ethnicity (92% white) and geography. A total of 74% (n = 595) of the patient population reported suffering from common mental health disorders, either depression only, anxiety only or both. Among the patient sample, 78% reported very poor to fair mental health

compared with 26% in the general population sample. The majority of respondents in the patient sample were female (61%) and 97% of the population were white. Half of the participants in each group completed ReQoL-10 and the other half completed ReQoL-20 **in their final formats**. A subset of each sample was asked to complete ReQoL-10 and ReQoL-20 approximately two weeks apart. The follow-up questionnaires were completed by 18% of participants from the patient (n = 141) and general population (n = 350) samples respectively.

Table 12 Characteristics of the online samples for reliability

		Patients (n = 800)		General population (n = 2,000)	
		Mean	**SD %**	**Mean N**	**SD %**
Age groups in years	18 to 24	25	3.1	223	11.2
	25 to 34	108	13.6	343	17.5
	35 to 44	147	18.4	334	16.7
	45 to 54	234	29.2	371	18.6
	55 to 64	273	34.1	296	14.8
	65 and over	13	1.6	433	21.7
Life satisfaction score	Score 0 to 10 (10 highest)	4.6	2.4	6.7	2.2
Gender	Male	311	38.9	927	46.4
	Female	489	61.1	1,073	53.6
Marital Status	Single	259	32.4	560	28.0
	Married / Partner	398	49.8	1,203	60.2
	Separated / Divorced	118	14.7	160	8.0
	Widowed	23	2.9	72	3.6
	Prefer not to say	2	0.2	5	0.2
Ethnicity	White	777	97.1	1,833	91.7
	Non white	23	2.9		
Degree	Yes	313	39.1	996	49.8
	No	487	60.9	1,004	50.2
Main activity	In employment or self-employment	332	41.5	1,063	53.2
	Retired	86	10.7	507	23.3
	Housework	95	11.9	162	8.1
	Student	19	2.4	101	5.1
	Unemployed	268	33.5	167	8.3
General physical health	Excellent	25	3.1	246	12.3
	Good	210	26.2	965	48.2
	Fair	303	37.9	566	28.3
	Poor	206	25.8	189	9.5
	Very poor	56	7.0	34	1.7
General mental health	Excellent	28	3.5	628	31.4
	Good	145	18.1	852	42.6
	Fair	357	44.6	407	20.3
	Poor	212	26.5	96	4.8
	Very poor	58	7.2	17	0.9

Sample (Validity): The sample used in the assessment of validity (Study 2 in Chapter 5 – Table 6) was recruited mainly from UK mental health secondary providers (n = 20) and general practices (n = 3), voluntary organisations (n = 2) across England, and additionally a trial cohort whose participants had consented to being approached for related research. Recruitment was mainly undertaken face-to-face by clinical studies officers in secondary care provider organisations, while recruitment from GP surgeries, trial cohort and voluntary organisation was by post. Booklets containing the 40-item set and a second measure (see below) were sent to each Trust to distribute to participants in clinics or post to them following an appointment.

From an estimated pool of 14,000 people, a total of 4,266 responses were returned, equating to response rates for secondary care provider organisations, GPs and the trial cohort of 51%, 20% and 53% respectively. In all, 46% of responses were recruited from outpatients' services, 27% from GPs, 13% from inpatients' services, 7% from charities and 6% from IAPT Services. The participants completed a 40-item set and therefore the results have been generated **from ReQoL-10 and ReQoL-20 items embedded in the 40-item set**. The sample was well distributed across groups defined by age, gender, marital status, main activity and ethnicity, with 39% reporting good to excellent mental health and 32% reporting poor to very poor general health. Participants presented from five main diagnostic groups: depression (1,166), bipolar (402), schizophrenia (394), anxiety (311), and personality disorder (233). The remaining groups were: other psychosis (123), eating disorders (102), substance abuse (102), behaviour (54) and sleep (39). Follow-up questionnaires were completed by 22% (953) approximately 4 to 12 weeks after the first administration of the baseline questionnaire. Life satisfaction was measured by the question used by the Office of National Statistics "Overall, how satisfied are you with your life nowadays?" on a scale of 0 (not at all satisfied) to 10 (completely satisfied) (34). The follow-up sample enjoyed higher life satisfaction as compared with baseline (6.9 vs. 5.3) and better health, with 19% reporting good to excellent mental health while 51% reporting poor to very poor general health.

7.1.2 Other measures

The SWEMWBS is a 7-item scale that captures positive affect of mental wellbeing in which each item is answered on a 1 to 5 Likert scale and (10). Transformed scores using Rasch are recommended for the SWEMWBS but in routine practice items are summed to produce a total score ranging from a minimum of 7 to a maximum of 35, with higher scores representing higher levels of mental wellbeing. The SWEMWBS was developed from the original 14-item version, which in turn was developed by a panel of experts and through the psychometric testing of an existing scale, Affectometer 2, which was developed for the general population in New Zealand (35).

The EQ-5D-5L is a generic preference-based measure that assesses health status on five dimensions: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression (36). There are five response options ranging from *extreme problems* to *no problems*. Utility scores are available for all 3,125 states defined by the measure with scores ranging from -.28 to .95, where zero is for states deemed to be as bad as being dead (37).

Clinical Outcomes in Routine Evalutation-10 (CORE-10) (38) is a 10-item shortened version of the 34-item Clinical Outcomes in Routine Evaluation-Outcome Measure (39, 40). The CORE-10 yields a simple psychometric structure and has broad coverage not only of depression and anxiety, but also of a wider spread of life functioning including general, social, and close relationships, as well as risk to self. Each item is scored on a 5-point scale from 0 *(not at all)* to 4 *(most of the time)*. The internal reliability (alpha) is .90 and the clinical cut-off score for general psychological distress is 11.0, with a maximum distress score of 40 (40).

The PHQ-9 is a self-administered version of the PRIME-MD diagnostic instrument for common mental disorders with each item being scored from 0 (not at all) to 3 (nearly every day), with a range from 0 to 27 (41). The PHQ-9 is the depression module, which scores each of the 9 diagnostic and

statistical manual of mental disorders (DSM-IV) criteria. The nine-item depression module from the PHQ-9 is well validated and widely used as a brief diagnostic and severity measure.

The GAD-7 is a self-administered seven-item measure of the severity of generalised anxiety disorder (GAD). It derives from 9 items that reflect all the symptom criteria for GAD and 4 items on the basis of a review of existing anxiety scales (42). Scores range between 0 and 21, classifying anxiety as minimal (0–4), mild (5–9), moderate (10–14) or severe (15–21). It has reported values of .92 for internal consistency and .83 for test–retest reliability (42).

7.1.3 Analyses

To aid direct comparisons, ReQoL-20 scores were halved so that both versions were within the range 0-40. ReQoL-10 scores were calculated if no more than one item had a missing response. The ReQoL-20 scores were calculated if no more than two items were missing. In both cases, the mean value of the other responses was used to impute the score for the missing item or items.

7.1.3.1 Reliability

We examined *the test-retest reliability* in two administrations of the ReQoL-10 and ReQoL-20 approximately two weeks apart in the online sample. Reliability was assessed by the intraclass coefficient (ICC), where an ICC > .70 would indicate very good test-retest reliability (43).

Internal reliability was assessed using Cronbach alpha to assess the extent to which the items were inter-related. Coefficients above .7 are acceptable, above .8 are good, and above .9 are excellent but above .94 suggest potential redundancy (43).

7.1.3.2 Construct validity

We examined two forms of construct validity: *convergent* and *known group validity*. For *convergent validity*, convergence between ReQoL and two other measures, SWEMWBS and CORE-10, was assessed using Pearson's product moment correlation coefficients and locally weighted scatterplot smoothing (LOWESS) techniques (44). LOWESS is a non-parametric regression technique designed to capture general patterns in the relationship between two measures without making assumptions about the actual relationship between the variables (44). It plots a line on a scatterplot on the central tendency between the two variables, and allows a visualization of the relationship between these variables across the score range. Strong correlations were expected between the ReQoL measures, SWEMWBS and CORE-10 as they reflect common mental health-related aspects of quality of life. Correlations are considered strong if scores are ≥0.7 (45).

7.1.3.3 Known Group Validity

Known group validity was examined in terms of whether the ReQoL measures were able to discriminate between people with a variety of specific conditions (i.e. depression, anxiety, schizophrenia, bipolar, personality disorder, and other conditions). For those with anxiety or depression, the known group validity was also assessed using GAD-7 and PHQ-9 cut-offs and by using CORE-10 clinical cut off points (where a score > 10 indicates clinical concerns). Whilst GAD-7 and PHQ-9 do not measure aspects of quality of life, they are thought to define broad groups expected to generate different quality of life scores. We also investigated known-group validity by using a self-reported global assessment of health and mental health. The five original categories were collapsed to binary categories of poor versus good health. Differences were quantified using standardised effect sizes (SES) across severity sub-groups calculated as the difference in mean scores between groups divided by the standard deviation of the milder of the two sub-groups. SES expressed as Cohen's *d* of 0.2 are normally considered small, 0.5 moderate, and 0.8 large (45).

7.1.3.4 Responsiveness

We measured responsiveness in two ways. First, we examined the numbers of people with either the lowest possible score or highest possible score since they impact on the ability of the ReQoL measures to detect deterioration or improvements respectively. Second, we used the sensitivity to apparent changes in quality of life. In the absence of an objective measure of change, we used the responses of people reporting mental health problems to a quality of life transition item that asked whether they thought their quality of life had stayed the same, improved (*somewhat* or *a lot*) or worsened (*somewhat* or *a lot*) since they last completed the questionnaire between 6 to 12 weeks ago. Responsiveness for ReQoL, SWEMWBS and EQ-5D was assessed using the standardised response mean (SRM) statistic, calculated by dividing the mean change on the measure by the standard deviation of the change. As for the other standardised effects sizes, SRMs of 0.2 are considered small, 0.5 moderate, and 0.8 large.

7.2 Results

7.2.1 Distribution of scores

The means and standard deviations (SDs) for ReQoL-10 and ReQoL-20 at baseline were 21.99 (SD = 10.26) and 21.63 (SD = 9.97) respectively. All items for both ReQoL measures were endorsed and the ReQoL scores populated the full range of the 0-40 and 0-80 scales. The overall score distributions were reasonably even across the score range, though there were some spikes and noticeably smaller numbers at the lower end of the scales.

As shown in Table 13 and Figures 9 and 10, the means (M) and standard deviations (SDs) for the three comparator measures were: EQ-5D (n = 1,592), M = .75 (SD = .25); summative and transformed SWEMWBS (n = 1103) scores, M = 23. 14 (SD = 6.80) and 21.71 (SD = 5.85) respectively; and CORE-10 (n = 216), M = 17.79 (SD = 10.94).

Table 13 Distribution of scores – ReQoL and other measures

	n	mean	standard deviation	completion rate %
ReQoL -10				
Baseline	4,037	21.99	10.26	95
Follow-up	953	24.18	10.08	
ReQoL -20 (scale 0 to 80)				95
Baseline	4,037	43.27	19.93	
Follow-up	953	48.56	19.57	
ReQoL -20 (scale 0 to 40)				95
Baseline	4,037	21.63	9.97	
Follow-up	953	24.28	9.78	
SWEMWBS total				
Baseline	1,103	23.14	6.80	95
Follow-up		24.35	6.43	
SWEMWBS rasch				
Baseline	1,103	21.71	5.85	95
Follow-up		22.64	5.66	
EQ-5D				
Baseline	1,592	0.75	0.25	98
Follow-up		0.78	0.22	
CORE-10				98
Baseline	216	17.79	10.94	
Follow-up	46	16.34	10.57	
PHQ-9				
Baseline	690	13.12	7.74	89
Follow-up		12.39	6.96	
GAD-7				
Baseline	554	6.24	5.18	96
Follow-up		12.08	7.44	

Note: ReQoL -10 in its embedded form of 40 items

Figure 9 Distribution of ReQoL-10 scores at baseline

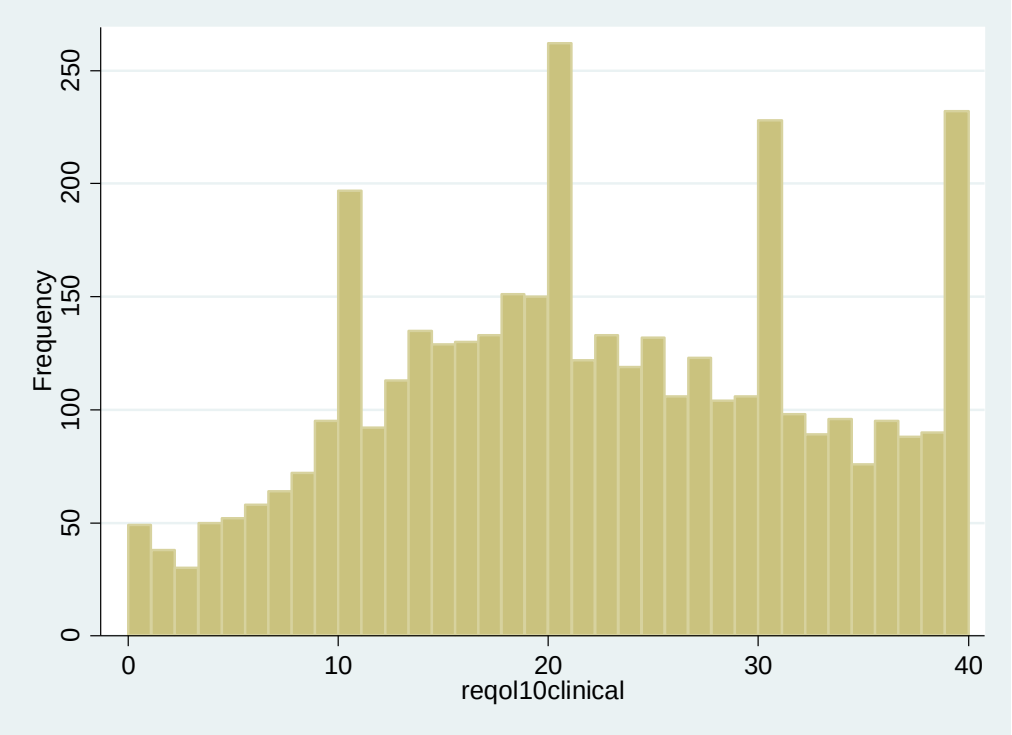
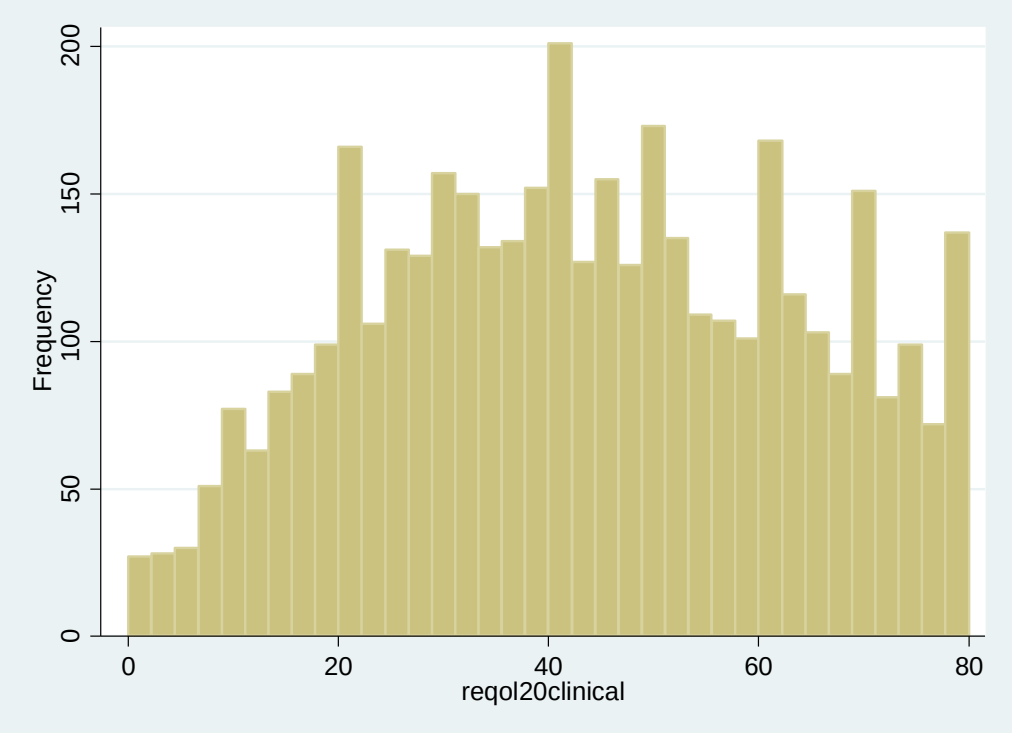


Figure 10 Distribution of ReQoL-20 scores at baseline on a scale 0 to 80



Missing data ranged between 3-4% for all mental health items of the ReQoL and scores could not be calculated for 5% of the sample where there were more than one and/or two items for the two measures. Imputation for missing data was performed for 5% and 11% of the sample to obtain ReQoL-10 and ReQoL-20 scores respectively. The completion rates for the comparative measures,EQ-5D, SWEMWBS and CORE-10, were above 95%.

7.2.2 Reliability

The ICC for the ReQoL-10 measure for both the general population sample (n = 488) and the patient sample (n = 279) reporting the same general mental health at both administrations was .85 ($p<.01$). For ReQoL-20 the ICC for the patient sample (n = 100) and the general population sample (n = 249) was .90 and .87 respectively. The ICC was also calculated at item level and ranged from .47 to .81. Cronbach alphas for the embedded ReQoL-10 and ReQoL-20 items in the main sample were .92 and .96 respectively. For the online samples, the equivalent alphas were .87 and .93 for ReQOL-10 and ReQoL-20 respectively.

7.2.3 Convergent validity

The correlations of both ReQoL measures with the summative score of the SWEMWBS and CORE-10 were above .80 across all four diagnostic groups and .90 or more for the pooled data set, suggesting a strong level of convergence (Table 14). The ReQoL-20 correlations were very similar to those of the ReQoL-10, though overall slightly higher (Figure 11). All correlations were significant ($p<.01$) and in the correct direction. LOWESS plots (Figures 12 and 13) show that the concordance appeared good between the ReQoL-10 and the SWEMWBS, and was better at the less severe end of the scale in both cases. The correlation between the ReQoL-10 and ReQoL-20 was .98.

Table 14 Convergence by condition of ReQoL measures with other measures

	All mental health conditions		Depression and Anxiety		Schizophrenia		Bipolar		Personality disorder	
	n	r	n	r	n	r	n	r	n	r
ReQoL-10 score with other measures										
SWEMWBS Total score	1,050	0.90	383	0.90	52	0.84	103	0.92	46	0.92
Rasch score		0.86		0.87		0.82		0.89		0.90
CORE-10	211	-0.88	55	-0.90	55	-0.76	25	-0.89	19	-0.89
ReQoL-20 score with other measures										
REQOL-10		0.98	1,470	0.98	517	0.96	402	0.97	233	0.97
SWEMWBS Total score	1,050	0.90	383	0.90	52	0.81	103	0.93	46	0.91
Rasch score		0.87		0.87		0.79		0.90		0.90
CORE-10	211	-0.93	55	-0.92	55	-0.87	25	-0.95	19	-0.96

All the correlation coefficients (r) are product moment correlations and significant at 1%

Figure 11 Lowess scatter plots between ReQoL-10 and ReQoL-20 (scale 0 to 40) at baseline

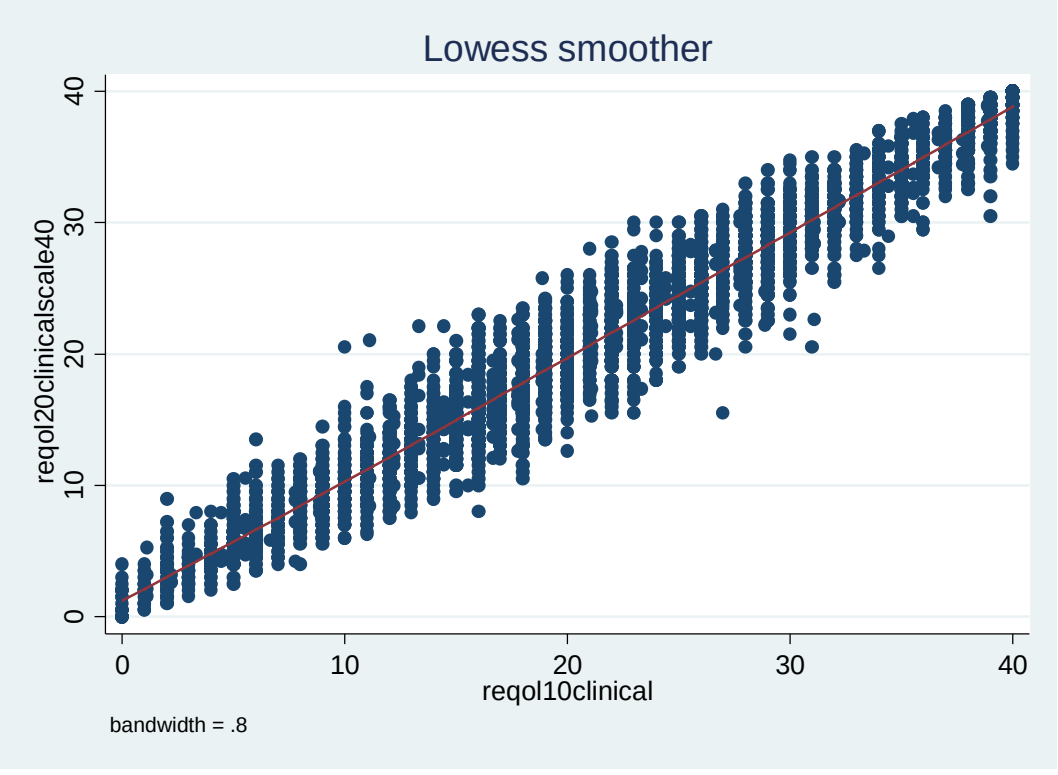


Figure 12 Lowess scatter plots between ReQoL-10 and SWEMWBS total score at baseline

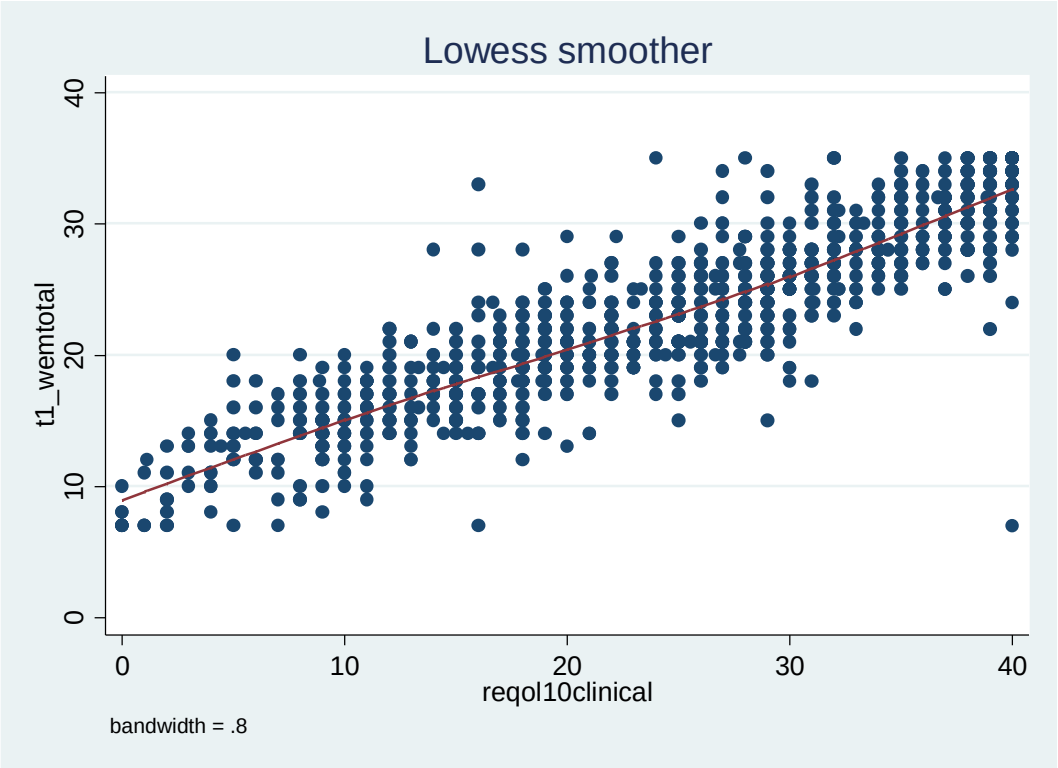


Figure 13 Lowess scatter plots between ReQoL-10 and SWEMWBS Rasch score at baseline

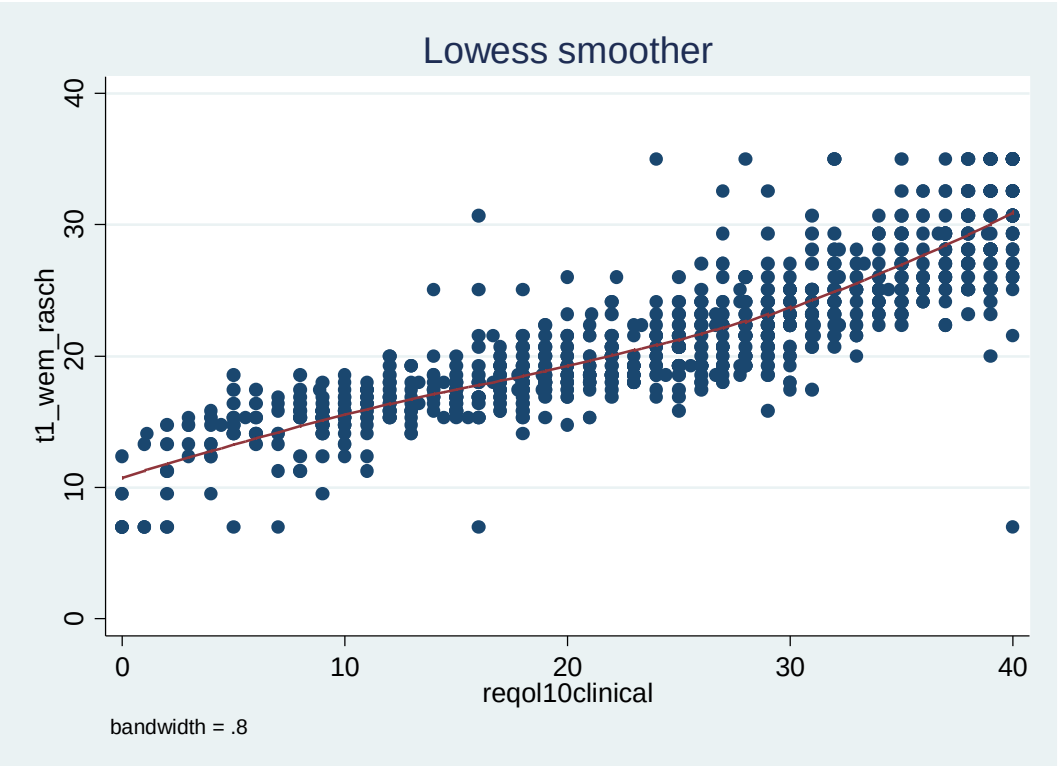
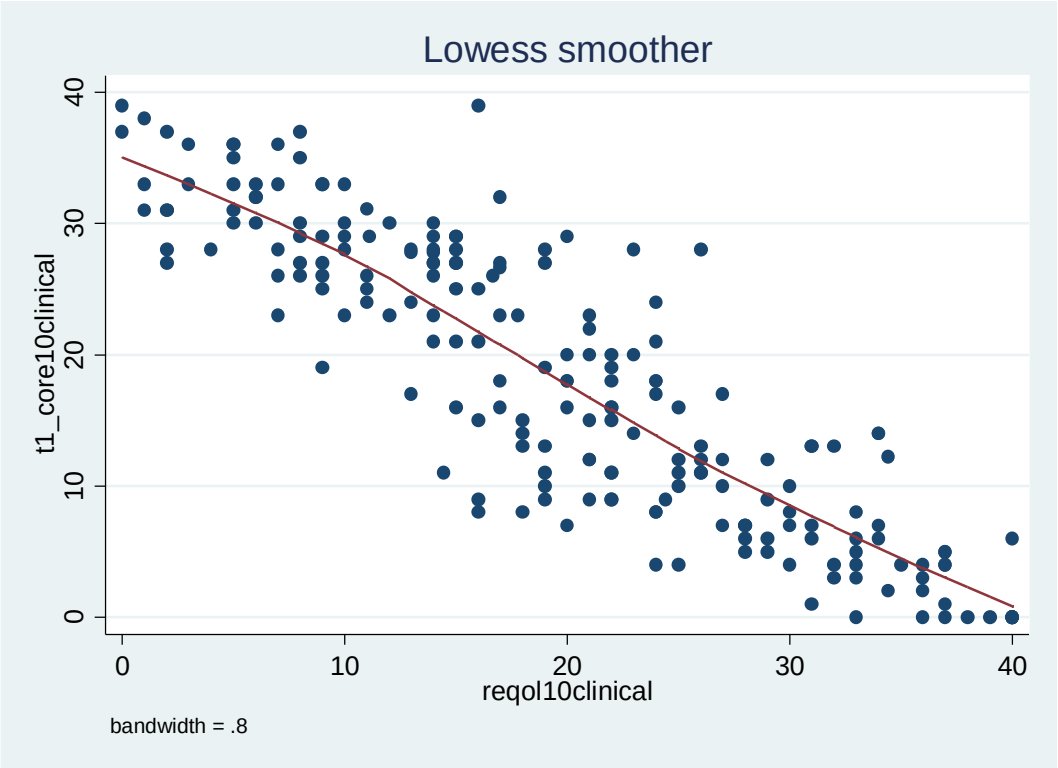


Figure 14 Lowess scatter plots between ReQoL-10 and CORE-10 at baseline



7.2.4 Known group validity

Table 15 presents the means, SDs, and SESs for the ReQoL-10 and ReQoL-20. The ReQoL scores for the online general population sample were significantly higher (i.e. healthier) compared with the six diagnostic groups of depression, anxiety, schizophrenia, bipolar, personality, and other diagnoses as broadly defined by ICD-10 codes (46). As shown in Table 15, the SESs show the differences were moderate for schizophrenia and other psychotic disorders and large for common mental health disorders, bipolar, personality and other mental health disorders. The SESs for ReQoL-20 were marginally larger than those for ReQoL-10. ReQoL scores distinguished between thresholds defined by the PHQ-9, GAD-7 and CORE-10. The largest SES was observed with CORE-10 cut-off and the lowest with GAD-7 score.

Table 15 Known group validity for ReQoL-10 and ReQoL-20

	ReQoL-10				ReQoL-20		
	N	Mean score	SD	SES	Mean score	SD	SES
General population	1,671	28.48	6.96	0.93	28.56	6.57	1.05
Patient population	4,037	21.99	10.26		21.63	9.97	
Using self-reported global assessment of mental health (good v poor)	2,633	26.56	8.40	1.68	26.17	8.10	1.72
	1,223	12.46	6.61		12.20	6.16	
Comparing general population with those who self-reported the following conditions							
Common mental health disorders	1,470	20.33	9.74	1.17	19.82	9.31	1.33
Schizophrenia and psychotic disorders	516	23.76	8.89	0.68	23.75	8.79	0.73
Bipolar	396	21.52	10.04	1.00	21.29	9.74	1.11
Personality Disorder	232	14.63	8.41	1.99	14.30	8.17	2.17
Other mental health disorders	252	18.58	9.73	1.42	17.87	8.99	1.63
Comparing by clinical cut-offs used in clinical practice							
PHQ-9 Clinical (score =>10)	419	15.73	7.53	1.70	15.23	7.08	1.85
Non-clinical	227	27.37	6.83		27.38	6.57	
GAD-7 Clinical (score =>8)	202	14.14	7.35	1.03	13.22	6.70	1.21
Non-clinical	318	23.02	8.63		23.10	8.18	
CORE-10 Clinical (score =>10)	150	15.08	8.02	2.37	14.85	7.50	2.96
Non-clinical	66	30.73	6.61		31.14	5.51	

Notes: p values are all <0.001

The known group differences analyses were repeated with the samples of participants who completed ReQoL and SWEMWBS and with those who completed ReQoL-10 and EQ-5D. Table 16 presents the means, SDs, and SESs for the ReQoL-10 when paired with either the SWEMWBS or the EQ-5D. The results comparing ReQoL-10 scores and SWEMWBS summative scores revealed higher SESs for ReQoL-10 in general (Table 16). When comparing ReQoL-10 scores with the transformed SWEMWBS scores, similar results were observed. The head-to-head comparison between ReQoL-10 and EQ-5D found the SESs to be markedly higher for ReQoL-10.

Table 16 Comparing known-group validity (SESs) of ReQoL-10, SWEMWBS and EQ-5D in same samples

	ReQoL-10 and SWEMWBS summative score			ReQoL-10 and EQ-5D		
		SESs			SESs	
	N	ReQoL-10	SWEMWBS	N	ReQoL-10	EQ-5D
General population v Patient population	1,007	0.56	0.48[1]	1,513	0.64	0.64[2]
Using self-reported global assessment of mental health (good v poor)	751 205	1.83	1.62	1,151 321	1.90	1.63
Comparing general population with those who self-reported the following conditions						
Common mental health disorders	371	0.78	0.66	530	0.92	0.68
Schizophrenia and psychotic disorders	52	0.76	0.63	190	0.56	0.44
Bipolar	98	0.91	0.75	97	0.77	0.64
Personality Disorder	46	2.09	1.89	59	1.83	1.15
Other mental health disorders	n = low			89	1.10	0.78

[1] This is the SWEMWBS transformed score as the norms for the general population norms are only provided for the transformed scores from: http://www2.warwick.ac.uk/fac/med/research/platform/wemwbs/researchers/interpretations/wemwbs_population_norms_in_health_survey_for_england_data_2011.pdf

[2] The EQ-5D norms have been provided by Devlin *et al*.

Tables A1 – A3 in the appendix provide more details on the comparison of SESs between ReQoL, SWEMWBS and EQ-5D.

7.2.4 Responsiveness

Scores improved for all instruments between administrations. For the 953 participants at follow-up, the mean and SDs for ReQoL-10 and ReQoL-20 were 24.18 (SD = 10.08) and 24.28 (SD = 9.78) respectively.

Table 17 Descriptive statistics of the ReQoL and other measures

	ReQoL-10	**ReQoL-20 (0 to 40 scale)**	**ReQoL-20 (0 to 80 scale)**	**SWEMWBS total score**	**SWEMWBS rasch score**	**EQ-5D**
Baseline scores						
N	4,037	4,037	4,037	1,103	1,103	1,592
Mean	21.99	21.63	43.27	23.14	21.71	0.75
Standard deviation	10.26	9.97	19.93	6.80	5.85	0.25
Follow up scores						
N	953	953	953	564	564	523
Mean	24.18	24.28	48.56	24.35	22.64	0.78
Standard deviation	10.08	9.78	19.57	6.43	5.66	0.22

Note: To compare the ReQoL-10 and ReQoL-20, ReQoL-20 has also been calculated on the scale of 0 to 40.

The proportions of responses at the worst scores were below 1% and less than 5% at the best level at both baseline and follow-up (Table 18).

Table 18 Floor and ceiling effects at baseline and follow-up

	% at worst score		% best score	
	T1	T2	T1	T2
ReQoL-10	0.72	0.63	3.77	4.6
ReQoL-20	0.30	0.32	1.49	1.9
SWEMWBS Total score	1.52	1.06	4.67	4.6
SWEMWBS Rasch score	1.52	1.06	4.67	4.6
EQ-5D	0.00	0.00	14.04	15.7

The SRMs for the ReQoL items were moderate for those reporting improvements in their health and those reporting deteriorations, and <.2 for those reporting their health had remained the same (Table 19).

Table 19 Responsiveness to change

	Health stayed the same			Health improved[1]			Health worsened[1]		
	n	mean (sd)	SRM	n	Mean (sd)	SRM	n	Mean (sd)	SRM
Comparing ReQoL and SWEMWBS									
ReQoL-10	272	-0.80(5.12)	-0.16	131	2.31 (6.04)	0.38	86	-3.61 (5.67)	-0.64
ReQoL-20	272	-0.49(4.54)	-0.11	131	2.47 (5.22)	0.47	86	-2.92 (5.33)	-0.55
SWEMWBS[2]	272	-0.15(4.05)	-0.04	131	1.60 (3.58)	0.45	86	-1.91 (3.82)	-0.5
SWEMWBS[3]	272	-0.15(3.86)	-0.04	131	1.27 (3.03)	0.42	86	-1.52 (3.19)	-0.48
Comparing ReQoL and EQ-5D									
ReQoL-10	123	0.50(6.59)	0.08	69	3.24 (8.35)	0.39	50	-3.82 (6.12)	-0.62
ReQoL-20	123	0.34 (5.94)	0.06	69	3.59 (7.65)	0.47	50	-3.47 (5.22)	-0.67
EQ-5D	123	0.17 (0.83)	0.08	69	0.82(0.18)	0.07	50	-0.04 (0.16)	-0.25

[1] Participants self-reported using a global assessment question at follow-up; and this combines the categories 'somewhat' and 'a lot'
SWEMWBS[2] – summative score SWEMWBS[3] – transformed Rasch score

Overall, SRMs between groups were similar in magnitude for SWEMWBS and both ReQoL versions. For patients reporting an improvement in health, the SWEMWBS SRMs were moderate in size and lay between those for the ReQoL-10 and ReQoL-20. The SRMs were marginally larger for the ReQoL instruments than the comparator measures in those who reported their health had worsened. The SRMs for EQ-5D were small according to Cohen's criteria and less than half those for the ReQoL versions in both groups of patients reporting change.

8 Discussion and conclusions

8.1 Summary of the development process

Following a four stage process, we developed two versions of the ReQoL measures: ReQoL-10 and ReQoL-20. Both measures contain a mixture of positively and negatively worded questions representing the polar aspects of the stages of recovery from a profound sense of loss and hopelessness to living a full and meaningful life (47). Both versions contain a physical question which does not contribute to the summative score. We obtained the themes for the measures from a systematic review and interviews with service users. The themes pertaining to the concepts of both recovery and quality of life were identified as: *activity* (meaningful and/or structured), *hope*, *belonging and relationships*, *self-perception*, *wellbeing*, *autonomy* and *physical health*. In Stage I of the process, we generated an initial selection of items from existing QoL and recovery measures and from transcripts from the qualitative interviews (6). We used a list of criteria (15) to shortlist items for the subsequent stage.

In Stage II, while examining the content validity of the ReQoL measure with 76 service users, we identified five key themes which helped us to exclude unacceptable items (*relevance of items*, *ease of response*, *item ambiguity*, *potentially distressing items* and *judgemental items*). We showed how the involvement of service users within this important methodological stage is paramount. To maximise the acceptability, validity and reliability, the findings from this research informed the development of the ReQoL measure at every stage; as a result of the feedback from service users, some items were reworded, while others were omitted from the later stages of the development of the measure. The findings were again used to inform the final item selection for the ReQoL measure alongside clinician input and psychometric assessment of item performance (Chapter 6).

In the psychometrics stage (Stage III) of the project, two studies were undertaken where over 6,500 service users have completed different item sets of the ReQoL items. Factor analysis confirmed the unidimensionality of the scale, with a bifactor model providing the best fit. Using graded IRT models, the items were modelled to identify potential misfitting items, item information functions and discrimination parameters. IRT scoring was compared with a simple summative scoring and, in the interest of ease, it was acceptable to use a simple summative score.

In Stage IV, we combined psychometric evidence with qualitative evidence from service users, clinicians and members from the governance groups to select the best items. This was done in a collaborative manner in a Scientific Group meeting with service users and clinicians present.

The reliability and validity results showed that all the response options were utilised and the whole measurement range was used. Results showed good internal reliability and test-retest reliability. Their construct validity was supported by strong convergence between the SWEMWBS and both ReQoL measures. The ReQoL measures are able to distinguish between the general and a patient population, those with four mental health conditions, and between those reporting good and poor mental health. Both ReQoL measures were able to detect changes when a change in mental health was reported. The SESs and SRMS for ReQoL-10 and ReQOL-20 were generally higher than SWEMWBS and markedly better than EQ-5D.

8.2 Discussion

A key hallmark of the ReQoL measures, though, is that they have been constructed with service users who determined the themes in the first instance (4, 6). Service users also tested the content validity of the items, selecting their preferred items and rephrasing items. In the psychometric stages of the project, service users completed the ReQoL item pools and their data was used to finalise the

two versions. Moreover, at each stage of the development process, service users were consulted and acted as decision-makers in the process. Clinicians were part of the study team and were also consulted at each stage of the project to ensure that the ReQoL items were clinically meaningful and useful to them. The involvement of service users is not only important for the face validity of the measures but also because of the long-standing recognition that their perspectives differ significantly from those of academics and clinicians (21). This development of the ReQoL illustrates the collaborative manner with which ReQoL was developed with service users and key stakeholders. The development process was transparent and inclusive, harnessing expertise from a range of contributors. In practical terms this will enhance acceptability, completion and usefulness. The content and face validity of ReQoL contribute to the latter being more embedded in mental health service users' lived experiences, as neither CORE-10 nor SWEMWBS had any input from service users (10, 38).

The ReQoL measures offer a number of important advantages over existing measures. They are the only ones known to the authors that have been built around the themes of recovery identified by Leamy and colleagues (3). In addition, the measures contain a mixture of positive and negative items which is a crucial element as people with mental health difficulties identified issues that both enhanced or depleted their quality of life. The presence of negative aspects increases the relevance of ReQoL as a PROM of recovery in mental health populations (48). Furthermore, an instrument such as SWEMWBS, which only contains positive items, may be masking deterioration and thereby inflating the positive impact of an intervention.

The ReQoL-10 measure was compared with EQ-5D because the latter is the recommended measure by NICE in the UK for use in economic evaluation of interventions (49). Previous work showed that EQ-5D was not suitable for use in the area of mental health (4, 13). The findings reported in this study show that the ReQoL provides a more sensitive and responsive measure than the EQ-5D. The

SRM for EQ-5D for participants reporting either an improvement or deterioration in health was .07 and .25 respectively, whereas the SRMs using ReQoL-10 were .38 and .64. This will have perverse implications when using EQ-5D to measure health benefits from a mental health intervention and may be disadvantageous in terms of the resources allocated to mental health services.

The findings of little difference between ReQoL-10 and ReQol-20 in terms of reliability and validity are not surprising given that the ReQoL-10 items are contained in the 20-item version. While the alpha of .96 for the ReQoL-20 suggests the presence of redundant items, all 20 items were retained in order to provide a fuller battery of items either for research studies or to provide a more rounded assessment in clinical settings and thereby enable clinicians to develop a greater understanding of service users' quality of life. ReQoL-20 can be used to support more in-depth conversations between clinicians and service users about which areas service users need most support with and to help clinicians and service users to understand progress during the intervention.

8.3 Caveats

A potential limitation in the validation results is that the properties of validity and responsiveness were assessed on the embedded ReQoL-10 and ReQoL-20 items contained in the 40–item set. However, we would not expect the results of the validation of the final measures to be different. In addition, participants would ideally be asked to complete the paired second measure chosen at random, but this was not practical. Instead all participants recruited from one organisation had the same paired measure, which may have introduced a hidden bias.

In the absence of a gold standard in this field, we had to rely on indirect methods of construct validity and responsiveness to provide evidence to support the properties of the measures. This should be seen alongside the qualitative evidence reported elsewhere on content and face validity.

Despite the large overall numbers, some self-identified presenting conditions had low numbers at baseline (e.g. less than 100 for substance abuse, behaviour and sleep) and all except depression had numbers below 100 at follow-up. The numbers fell even more when categorised by change in health status. As a result, in the head to head comparison, we could not investigate SRM by conditions. Further, we have used crude measures of known group validity as they were the only ones that could feasibly be collected during the study. The crude groupings need further refinement to be able to build on the evidence presented.

8.4 Ongoing work

To aid interpretation of the ReQoL scores, norms need to be developed and minimal important differences estimated. To use ReQoL for economic evaluation preference weights will be estimated, allowing the computation of quality adjusted life years for use in cost-effectiveness analysis. Further validation in its final formats is warranted. For the ReQoL measures to be used more widely, it is important to examine them across different cultures and across different countries.

8.5 Conclusion

By virtue of its brevity, the ReQoL-10 measure only takes a few minutes to complete, making it convenient for use in routine practice. The items are short, clear and easy to answer. The ReQoL-20, by offering more granular information, might be used in an initial or discharge assessment and requires approximately five minutes to complete. It might also be suitable to include in clinical studies (e.g. randomised controlled trials). ReQoL should offer significant advantages compared with generic measures like the EQ-5D and SWEMWBS that were not developed in conjunction with mental health service users, as well as measures based on symptoms from one disorder, like the

PHQ-9, which is commonly used in clinics but does not reflect the broader concerns of many service users beyond depressive symptomatology.

Box 1 The strengths of ReQoL

- ✓ Collaboratively developed with service users and clinicians who were central to the research, as advisors, researchers, and participants
- ✓ Tested by 6000+ service users and is psychometrically validated
- ✓ Suitable for the whole spectrum of mental health conditions, from common mental health disorders through to very severe ones
- ✓ Free to use for NHS and publicly funded health care and research
- ✓ Can be incorporated into patient information systems
- ✓ Can capture service users' perspectives
- ✓ Can be integrated into care planning and used to inform care decisions with service users as participants in decision-making processes
- ✓ Can be used as a therapeutic tool to guide conversations and help focus sessions
- ✓ Can be used to provide on-going feedback of progress
- ✓ Suitable for ages 16+ and for people with different cultural backgrounds
- ✓ Easy to complete as it is short and simple
- ✓ Scores can be easily calculated and interpreted

Further details, sample copies, scoring guides and requests for permissions to use the ReQoL measures are available from the Clinical Outcomes team at Oxford University Innovation Ltd at: **http://innovation.ox.ac.uk/outcome-measures/recovering-quality-life-reqol-questionnaire/.**

9 References

1. Department of Health. Mental Health Clustering Booklet (2013/14). 2013.
2. Boardman J SM, Shepherd G. Assessing recovery: seeking agreement about the key domains. Report for the Department of Health. 2013.
3. Leamy M, Bird V, Le Boutillier C, Williams J, Slade M. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. The British Journal of Psychiatry. 2011; 199 (6): 445-52.
4. Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, *et al*. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technology Assessessment. 2014; 18 (34).
5. Connell J, Brazier J, O'Cathain A, Lloyd-Jones M, Paisley S. Quality of life of people with mental health problems: a synthesis of qualitative research. Health and Quality of Life Outcomes. 2012; 10 (1): 138.
6. Connell J, O'Cathain A, Brazier J. Measuring quality of life in mental health: Are we asking the right questions? Soc Sci Med. 2014; 120: 12-20.
7. Department of Health. Mental Health Payment by Results: Quality and Outcomes Framework Report. 2013.
8. CPPP. Results for Pilot of Short Version of the Warwick and Edinburgh Mental Well Being Scale (SWEMWBS). Care Pathways and Packages Project. 2014.
9. Kammann R FR. Affectometer 2: a scale to measure current level of general happiness. Aust J Psychol. 1983: 259-65.
10. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, *et al*. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health and Quality of life Outcomes. 2007; 5 (1): 1.
11. Keyes CL. Mental illness and/or mental health? Investigating axioms of the complete state model of health. J Consult Clin Psychol. 2005.
12. Newsom JT, Rook KS, Nishishiba M, Sorkin DH, Mahan TL. Understanding the relative importance of positive and negative social exchanges: Examining specific domains and appraisals. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2005; 60 (6): 304-12.
13. Brazier J. Is the EQ–5D fit for purpose in mental health? The British Journal of Psychiatry. 2010; 197 (5): 348-9.
14. Saarni SI, Viertiö S, Perälä J, Koskinen S, Lönnqvist J, Suvisaari J. Quality of life of people with schizophrenia, bipolar disorder and other psychotic disorders. The British Journal of Psychiatry. 2010;197(5):386-94.
15. Streiner DLGR, Norman. Health measurement scales: a practical guide to their development and use. Oxford Medical Publications. 1989.
16. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. 1998.
17. Holden RR. Face validity. Corsini Encyclopedia of Psychology. 2010.
18. Staniszewska S, Haywood KL, Brett J, Tutton L. Patient and public involvement in patient-reported outcome measures. The Patient: Patient-Centered Outcomes Research. 2012; 5 (2): 79-87.
19. US Department of Health and Human Services FDA Center for Drug Evaluation and Research, US Department of Health and Human Services FDA Center for Biologics Evaluation and Research, US Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in

medical product development to support labeling claims: draft guidance. Health and Quality of Life Outcomes. 2006; 4: 1-20.

20. Perkins R. What constitutes success? Br J Psychiatry. 2001; 179 (1).
21. Rose D, Evans J, Sweeney A, Wykes T. A model for developing outcome measures from the perspectives of mental health service users. Int Rev Psychiatry. 2011; 23 (1): 41–6.
22. Conway K, Acquadro C, Patrick DL. Usefulness of translatability assessment: results from a retrospective study. Qual Life Res. 2014; 23 (4): 1199-210.
23. Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychological methods. 2004; 9 (4): 466.
24. Muthen LK MB. Mplus User's Guide (7th edn). Muthen & Muthen. 1998–2015.
25. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement. 1969.
26. Orlando M. TO. New item fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24: 50-64.
27. Cai L, Thissen D, du Toit S. IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International. 2011.
28. Zumbo BD, editor. A handbook on the theory and methods of differential item functioning (DIF). 1999.
29. Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika. 1991; 78 (3): 691-2.
30. Newson R. Confidence intervals for rank statistics: Somers' D and extensions. Stata Journal. 2006; 6 (3): 309.
31. StataCorp. 2015. College Station TSL. Stata Statistical Software: Release 14. 2015.
32. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. 1981; 46 (4): 443-59.
33. Uebersax JS. Statistical modeling of expert ratings on medical treatment appropriateness. Journal of the American Statistical Association. 1993; 88 (422): 421-7.
34. Tinkler L, Hicks S. Measuring subjective well-being. London: Office for National Statistics. 2011: 443-55.
35. Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. Health and Quality of Life Outcomes. 2009; 7 (1): 1.
36. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, *et al*. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011; 20 (10): 1727-36.
37. Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. London: Office of Health Economics; 2016.
38. Barkham M, Bewick B, Mullin T, Gilbody S, Connell J, Cahill J, *et al*. The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. Counselling and Psychotherapy Research. 2013; 13 (1): 3-13.
39. Barkham M, Margison F, Leach C, Lucock M, Mellor-Clark J, Evans C, *et al*. Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. J Consult Clin Psychol. 2001; 69 (2): 184.
40. Evans C, Connell J, Barkham M, Margison F, McGrath G, Mellor-Clark J, *et al*. Towards a standardised brief outcome measure: psychometric properties and utility of the CORE—OM. The British Journal of Psychiatry. 2002; 180 (1): 51-60.
41. Kroenke K, Spitzer RL, Williams JB. The Phq-9. J Gen Intern Med. 2001; 16 (9): 606-13.
42. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006; 166 (10): 1092-7.

43. Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes: John Wiley & Sons; 2013.
44. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association. 1979; 74 (368): 829-36.
45. Cohen J. Statistical power analysis for the behavior science. Lawrance Eribaum Association. 1988.
46. World Health Organization (WHO). International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) Version for 2010. 2010. [Available from: http://apps.who.int/classifications/icd10/browse/2010/en#/F30-F39.
47. Andresen R, Oades L, Caputi P. The experience of recovery from schizophrenia: towards an empirically validated stage model. Aust N Z J Psychiatry. 2003; 37 (5): 586-94.
48. Crawford MJ, Robotham D, Thana L, Patterson S, Weaver T, Barber R, *et al*. Selecting outcome measures in mental health: the views of service users. Journal of Mental Health. 2011; 20 (4): 336-46.
49. National Institute for Health and Care Excellence (NICE). Guide to the Methods of Technology Appraisal 2013.

10 Appendices

Table A1: Known group validity: comparing ReQoL-10 and EQ-5D

	ReQoL-10				EQ-5D			
	n	mean(sd)	p value	ES	n	mean(sd)	p value	ES
General population v patient population	1,671 1,513	28.48 (6.96) 24.02 (10.04)	<0.001	0.64	996 1,513	0.88 (0.21) 0.75 (0.25)	<0.001	0.59
Comparing general population and the main disease areas								
Common mental health disorders	530	22.10 (9.61)	<0.001	0.92	530	0.73 (0.25)	<0.001	0.68
Psychotic disorders	190	24.61 (9.40)	<0.001	0.56	190	0.78 (0.23)	<0.001	0.44
Bipolar	97	23.13 (9.47)	<0.001	0.77	97	0.74 (0.26)	<0.001	0.64
Personality disorder	59	15.71 (8.47)	<0.001	1.83	59	0.63 (0.27)	<0.001	1.15
Other MH disorders	89	20.82 (9.96)	<0.001	1.10	89	0.71 (0.26)	<0.001	0.78
Using self-reported global assessment of health (Good versus Poor)	893 572	27.62 (8.90) 18.47 (9.12)	<0.001	1.03	893 572	0.87 (0.13) 0.57 (0.27)	<0.001	2.31
Using self-reported global assessment of **mental** health (Good versus Poor)	1,151 321	27.44 (8.12) 12.00 (6.39)	<0.001	1.90	1,151 321	0.82 (0.19) 0.51 (0.28)	<0.001	1.63

Table A2: Known group validity: comparing ReQoL-10 and SWEMWBS transformed (Rasch) score

	ReQoL-10				SWEMWBS transformed			
	n	mean(SD)	ES	p value	n	mean(SD)	ES	p value
General population v patient population	1,671 1,007	28.48 (6.96) 24.61 (10.62)	0.56	<0.001	7,196 1,007	23.61 (3.9) 21.73 (5.86)	0.48	<0.001
Comparing general population and the main disease areas								
Common mental health disorders	371	23.38 (10.09)	0.78	<0.001	371	20.73 (4.71)	0.74	<0.001
Psychotic disorders	52	23.20 (9.52)	0.76	<0.001	52	21.15 (5.59)	0.63	<0.001
Bipolar	98	22.12 (10.31)	0.91	<0.001	98	20.70 (5.83)	0.75	<0.001
Personality disorder	46	13.93 (8.41)	2.09	<0.001	46	16.45 (4.70)	1.84	<0.001
Other disorders	n low							
Using self-reported global assessment of health (Good versus Poor)	583 379	28.55 (9.34) 18.54 (9.44)	1.07	<0.001	583 379	23.71 (5.67) 18.62 (4.46)	0.90	<0.001
Using self-reported global assessment of **mental** health (Good versus Poor)	751 205	28.19 (8.60) 12.42 (6.91)	1.83	<0.001	751 205	23.38 (5.32) 16.25 (3.48)	1.34	<0.001

Table A3: Known group validity: comparing ReQoL-10 and SWEMWBS total (summative) score

	ReQoL-10				SWEMWBS total			
	n	mean(SD)	ES	p value	n	mean(SD)	ES	p value
General population v patient population	1,671 1,007	28.48 (6.96) 24.61 (10.62)	0.56	<0.001	7,196 1,007	25.30 (4.72) 23.25 (6.81)	0.43	<0.001
Comparing general population and the main disease areas								
Common mental health disorders	371	23.38 (10.09)	0.78	<0.001	371	22.19 (5.89)	0.66	<0.001
Psychotic disorders	52	23.20 (9.52)	0.76	<0.001	52	22.31 (6.51)	0.63	<0.001
Bipolar	98	22.12 (10.31)	0.91	<0.001	98	21.74 (6.83)	0.75	<0.001
Personality disorder	46	13.93 (8.41)	2.09	<0.001	46	16.4 (5.82)	1.89	<0.001
Other disorders	n low							
Using self-reported global assessment of health (Good versus Poor)	583 379	28.55 (9.34) 18.54 (9.44)	1.07	<0.001	583 379	25.56 (6.14) 19.44 (5.87)	1.00	<0.001
Using self-reported global assessment of **mental** health (Good versus Poor)	751 205	28.19 (8.60) 12.42 (6.91)	1.83	<0.001	751 205	25.29 (5.74) 16.00 (4.60)	1.62	<0.001

All the correlation coefficients are significant at 1%.