Towards a more inclusive approach to measure development in mental health populations: integrating qualitative and quantitative data in the Recovering Quality of Life (ReQoL) measure

Authors: Keetharuth, A., Taylor Buck, E., Acquadro, C., Conway, K., Connell, J., Barkham, M., Carlton, J., Ricketts, T., Barber, R., and Brazier, J.

Corresponding author: Anju Keetharuth
SchARR, University of Sheffield,
Regent Court, 30 Regent Street, Sheffield, S1 4DA
Tel: (+44) (0)114 222 0884
Fax: (+44) (0)114 222 0710
Email: d.keetharuth@sheffield.ac.uk

No. 18.02
Disclaimer:
This series is intended to promote discussion and to provide information about work in progress. The views expressed in this series are those of the authors. Comments are welcome, and should be sent to the corresponding author.

This paper is also hosted on the White Rose Repository: http://eprints.whiterose.ac.uk/
TITLE: Towards a more inclusive approach to measure development in mental health populations: integrating qualitative and quantitative data in the Recovering Quality of Life (ReQoL) measure

Authors: Anju Devianee Keetharuth¹, Elizabeth Taylor Buck², Catherine Acquadro³, Katrin Conway⁴, Janice Connell⁵, Michael Barkham⁶, Jill Carlton⁷, Thomas Ricketts⁸, Rosemary Barber⁹ and John Brazier¹⁰

Correspondence: d.keetharuth@sheffield.ac.uk Tel: +44 1142220884

¹ School of Health and Related Research, University of Sheffield, S14DA United Kingdom; d.keetharuth@sheffield.ac.uk
² School of Health and Related Research, University of Sheffield, S14DA United Kingdom; e.taylor-buck@sheffield.ac.uk
³ Mapi Research Trust, 27 Rue de la Villette, 69003 Lyon, France; cacquadro@mapigroup.com
⁴ Mapi Research Trust, 27 Rue de la Villette, 69003 Lyon, France; kconway@mapigroup.com
⁵ School of Health and Related Research, University of Sheffield, S14DA United Kingdom; j.connell@sheffield.ac.uk
⁶ Centre for Psychological Services Research, University of Sheffield, S102TN, United Kingdom; m.barkham@sheffield.ac.uk
⁷ School of Health and Related Research, University of Sheffield, S14DA United Kingdom; j.carlton@sheffield.ac.uk
⁸ School of Health and Related Research, University of Sheffield, S14DA United Kingdom; t.ricketts@sheffield.ac.uk
⁹ School of Health and Related Research, University of Sheffield, S14DA United Kingdom; rosemary.barber@sheffield.ac.uk
¹⁰ School of Health and Related Research, University of Sheffield, S14DA United Kingdom; j.e.brazier@sheffield.ac.uk
Abstract:

With the increasing prevalence of mental health problems, there is a need for a short patient reported outcome measure to assess the quality of life outcomes for individuals living with mental health conditions. While it is important to treat symptoms, there is growing recognition that in order to help people with mental health problems lead meaningful and fulfilling lives, it is crucial to capture the impact of their conditions on wider aspects of their social lives. We constructed two versions of the Recovering Quality of Life (ReQoL) measure for use in routine settings and clinical trials from a larger pool of items by combining qualitative and quantitative evidence covering six domains. Qualitative evidence was gathered through interviews and focus groups with service users and clinicians and a translatable assessment. Psychometric evidence was obtained from confirmatory factor models and item response theory analyses. A novel approach based on a traffic light pictorial format was developed to present qualitative and quantitative evidence to a group of service users, clinicians and researchers to help make the final selection. This work provides a pragmatic yet rigorous approach to combining qualitative and quantitative evidence to ensure that ReQoL is psychometrically robust and has high relevance to service users and clinicians.

Keywords: public mental health; quality of life; measuring outcomes; mixed methods; recovery
1. Introduction

With the increasing prevalence of mental health problems [1], there is a need for a short patient reported outcome measure (PROM) to assess the quality of life outcomes for individuals living with mental health conditions. While it is important to treat symptoms, there is a growing recognition of the value of leading a meaningful life and the need to capture the impact of conditions on this, even in the presence of symptoms. It is also known that many people with mental health conditions are able to lead fully ‘functional lives’ despite the presence of symptoms. Most importantly this ‘recovery’ process “is best judged by the person living with the experience” [2](p. 3). However, to capture the recovery process requires some form of measurement, and it is the form that such measurement takes that is the focus of the present article.

To date, recovery has largely been viewed as a reduction in an individual’s symptoms, with many measures focusing on single diagnostic entities such as depression (e.g., Personal Health Questionnaire; PHQ-9, [3]) or anxiety (e.g., Generalized Anxiety Disorder; GAD-7, [4]). The use of such mono-symptomatic measures defines recovery in a restricted and diagnostic fashion, thereby isolating the individual from their social context. Such measures also define the individual solely in terms of specific symptoms rather than more fundamental human aspects that relate to a person’s sense of self and their relationship within a social world. Measurement needs to incorporate multiple aspects of the individual of which symptoms are just one component. The challenge for measure developers is combining a more inclusive approach to defining recovery with a robust psychometric approach that makes the resulting instrument a valid and reliable measure of recovery that has meaning and relevance for the individual.

The process of constructing outcome measures can be seen as occurring at specific points along a continuum. At one end they can be constructed wholly by researchers and/or mental health professional experts with no or very limited involvement of service users, despite the latter being the
focus of the outcome measure. At the other end of the continuum, some advocate that service users should be the sole researchers [5]. In the middle of this spectrum is the principle of co-production involving service users in the development process together with researchers and clinicians. However, the scope and extent of service user involvement can vary significantly [6]. In this article, we report on the co-production of a service user-valued recovery measure named Recovering Quality of Life (ReQoL) from which we derived two versions (ReQoL-10 and ReQoL-20) for use with mental health service users [7]. The term user-valued is employed to convey the fact that the measures reflect the experiences of most service users and that they can relate to the themes and questions in the measure in the context of their lived experiences [8].

The ReQoL measures were developed in four stages as shown in Figure 1. The theoretical background for the measure identified six mental health themes (activity; belonging and relationship; choice and control; hope; self-perception; well-being) and one physical health theme [9-11]. In Stage I, items were generated under each theme using those from existing quality of life and recovery measures; phrases from the interview transcripts used to identify the themes [11]; and items identified by the research team. These items, 1597 in all, were sifted using an adapted criteria list [7, 12] to arrive at 87 items. In Stage II, these items were presented in turn to working age adult service users and younger service users to consider their appropriateness. Qualitative data on the items were also gathered on the 61-item set through a translatability assessment (Figure 1). In the third stage of the project, psychometric analyses were carried out in two separate studies recruiting 2262 and 4266 participants respectively. The qualitative evidence was integrated with the quantitative data to produce the final measures in Stage IV. In terms of governance of the project, the members of the Stakeholders Group (n = 33), the Advisory Group (n = 32), and the Expert Users Group (n = 6) were asked to comment at each stage of the project. The members of the Psychometrics Group (n = 6) provided specialist advice on the quantitative studies. In addition, the six Expert Service Users were also members of the Scientific Group (n = 18) which was the decision-making group.
Figure 1 Stages in the development of the ReQoL

Theoretical framework of the ReQoL based on a systematic review and qualitative interviews identified the following themes: activity; belonging and relationships; choice, control and autonomy; hope; self-perception; wellbeing and; physical health.

Stage I
Generation of 1597 items from: manuscripts, existing measures and coining new ones

Stage II
Generation of Qualitative evidence from:
  a. Interviews and focus groups with adults (n = 59) and younger service users (n = 17), clinicians (n = 35)
  b. Translatability assessment (wave 1 n = 87 items) (wave 2 n = 11 items)

Stage III
Generation of Quantitative psychometric evidence from service users in Study 1 (n=2262) and Study 2 (n=4266)

Stage IV
Combining the evidence from Stages II and III to construct the final ReQoL measures using a novel more socially and methodologically inclusive approach (Focus of this paper)

ReQol-10
Containing 10 mental health items and 1 physical health item

ReQol-20
Containing 20 mental health items and 1 physical health item
The aim of this paper is to present a novel approach of combining qualitative and quantitative evidence in the construction of an outcome measure that has high face and content validity with service users and clinicians consistent with the view of a broader definition of recovery.

2. Materials and methods

2.1 Sources of qualitative evidence

2.1.1 Qualitative data from service users of working age

Fifty-nine service users were recruited from four NHS Trusts and they were presented with a subset of the 87 items to avoid respondent burden. However, we ensured that all the items were presented to service users. This sample is discussed in detail elsewhere [13] but in summary, the sample included people experiencing a range of conditions including depression, anxiety, schizophrenia, and personality disorders. The mean age was 42 years and 63% were female. Service users were asked to: provide their feedback on the suitability of the items; rephrase items where necessary and; choose their preferred items where there were several covering a similar sub-theme.

2.1.2 Qualitative data from service users aged 16 to 18 years of age

The original interviews to identify the themes of the ReQoL measures excluded those aged below 18 and the youngest interviewee was aged 19. As the ReQoL measures would be used in a population as young as 16 years of age, it was essential to ensure that the themes and items resonated with this younger group. Seventeen participants were recruited from two Child and Adolescent Mental Health Services. All participants were students, and two also worked part-time. They were presented with the item set of 61 items but were also asked if any themes were missing. They were consulted on a reduced item set because their interviews happened later in the process and items considered most unsuitable by the adult group were already omitted.
2.1.3 Qualitative data from clinicians

Clinicians were consulted throughout the development process but it was deemed necessary to obtain specific feedback at the end of each stage from clinicians who were eventually going to be using the ReQoL measures with their patients. Thirty-five clinicians were interviewed in focus groups from six NHS trusts. They were asked to provide feedback on the relevance and usefulness of the items from their viewpoint and whether they already asked the information conveyed by the items as part of their routine consultations.

2.1.4 Translatability assessment

A translatability assessment (TA) was carried out by two linguists on the first item set resulting from stage I (i.e., Generation of candidate items - \( n = 87 \) (wave 1), and then, on the set of new items generated during stage II (i.e., Content validation - \( n = 11 \) (wave 2). TA is the review of its source text preferably during the development stage in order to determine its suitability for future translations in multilingual studies. The goal of TA is to facilitate future translations and use of the measure in global studies by: 1) assessing the interpretation of the each item’s underlying concept; 2) identifying and categorising potential translation issues in the source text (either cultural, semantic, idiomatic, syntactic or structural) and; 3) providing alternative choices of wordings on which translations can be based and/or recommendations of how to modify the source text (i.e., reformulation or deletion of item) so that future translations are conceptually and culturally appropriate for the target populations [14].

2.2 Analysis of qualitative data

A pragmatic approach was adopted in the analyses where comments made by each participant were added in a column next to the item in an Excel sheet. Taking into consideration all the comments for each item, a traffic light system was used to highlight items with predominantly negative comments (red), neutral and mixed items (amber), and positive comments (green).
2.3 Quantitative data

2.3.1. Study 1

In Study 1, 2262 service users were recruited from NHS secondary care providers, GP surgeries, a trial cohort and voluntary organisations and they were asked to complete a 61-item set. The mean age of 48 years (range 16 to 97) and participants (58% females) presented with a wide range of conditions including depression, anxiety, schizophrenia, bipolar and personality disorders, etc.

2.3.2. Study 2

In Study 2, service users were recruited from similar organisations as in the previous study and they were asked to complete the 40-item set (see Section 3.3.1 for the results of reducing the number of items from 61 to 40). The mean age was 47 years (range 16 to 98) and services users presented with a wide range of conditions. The samples and results for both studies are described in detail elsewhere [7, 15].

2.3.3 Analyses

Similar analyses were carried out in both studies. First, a confirmatory factor analysis (CFA) was undertaken to test whether the six themes adequately represented the structure of the data. Then an Exploratory factor analysis (EFA) was carried out followed by a CFA where a two-factor model and a bi-factor model were estimated. Item response theory (IRT) analyses were carried out using Graded Response Models (GRM) to determine the psychometric properties of the items. Classical psychometric analyses were also undertaken, namely responsiveness analyses in Study 2. The Psychometrics Group established a set of criteria to choose the items with the best properties (Table 1).
Table 1 Criteria to assess psychometric evidence

<table>
<thead>
<tr>
<th>Analyses</th>
<th>Guidance/judgement/rule</th>
<th>Determines exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Missing data at item level</td>
<td>If any item has ≥5% missing data, this item should be dropped</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>Lay description:</strong> Respondents do not complete an item maybe because they find it difficult, ambiguous or simply choose not to answer it.</td>
<td></td>
</tr>
<tr>
<td>2. Factor analyses</td>
<td>Identify items with high residual correlations (&gt;0.1)</td>
<td>Select one item</td>
</tr>
<tr>
<td></td>
<td><strong>Lay description:</strong> Pairs of items are identified as being very similar to each other in terms of what they are measuring and therefore one of the two would suffice for the measure.</td>
<td></td>
</tr>
<tr>
<td>3. IRT – Misfitting items</td>
<td>Identify items with p values &lt; 0.05 in at least 3 datasets (the sample was randomly distributed in 4 datasets of about 1000 observations in each).</td>
<td>No. Acknowledge the misfit but retain item in the item pool.</td>
</tr>
<tr>
<td></td>
<td><strong>Lay description:</strong> A mathematical model using all the items is used to explain whether all the items are measuring quality of life (QoL). Items can be identified as misfitting if they are not contributing to measuring QoL. This issue can arise not because of the item but because of the respondents. Therefore the item can be tested in a different sample before deciding whether to drop it.</td>
<td></td>
</tr>
<tr>
<td>4. IRT – Information functions</td>
<td>Ensure that items cover the whole measurement range (i.e., intensity) by choosing items to balance maximization of information over the total range and content validity (including items from all themes)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td><strong>Lay description:</strong> Highly discriminating items provide great information but over a small range of QoL and less discriminating items provide less information but over a wider range of QoL. The graphs for each item therefore tells us how much information an item is contributing to the scale and also to what portion of the score range (that those with very low or high QoL)</td>
<td></td>
</tr>
<tr>
<td>5. Differential Item Functioning (DIF)</td>
<td>Exclude items with DIF (age, ethnicity, gender, mental health condition)</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>Lay description:</strong> An item is said to display DIF if people with same QoL respond differently to the item because of other characteristic (e.g. by virtue of being female, or belonging to a particular socio-economic group). The item is picking these characteristics up and therefore not correctly representing the true QoL.</td>
<td></td>
</tr>
<tr>
<td>6. Sensitivity to change</td>
<td>Ensure selected items show change in response over time</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>Lay description:</strong> Given that ReQoL would be used routinely over the course of treatment, the items need to be able to register a change in the respondent’s QoL if there is one.</td>
<td></td>
</tr>
</tbody>
</table>
2.4 Combining qualitative and quantitative evidence

2.4.1 Study 1

The aim of this stage was to further reduce the number of items to around 40 to reduce respondent burden in the subsequent quantitative study. The qualitative and quantitative information were presented in tabular form (see Section 3.3.1). The psychometrics evidence for each of the 61 items was provided in terms of: the mean IRT value (how severe the item is), the area under the information curve (the amount of information the item is providing), highly correlated pairs of items and whether the item is misfitting. At the Scientific Group meeting chaired by the Chief Investigator (JB), various researchers synthesised the psychometrics evidence (AK), qualitative evidence - adult service users (JCo) and qualitative evidence younger service users (ETB) for each item, translatability assessment (AK). Each item was discussed in their themes and the ‘worse’ items were identified for deletion.

2.4.2 Study 2

With a view to simplifying the information required for members of the Scientific Group to make a decision for the final item selection, a more intuitive pictorial representation of the evidence was designed. The psychometrics evidence for each item was also colour coded using the traffic light system by a researcher (AK) based on the criteria in Table 1. The various terms used in Table 1 were explained in lay terms to ensure proper understanding. The qualitative evidence was colour coded by the qualitative researchers (JCo, ETB and TR). The evidence was then presented in diagrammatic format (see Section 3.3.2). A similar diagram was prepared for each item under a theme and sub-theme. In cases where two items from different themes were highly correlated as per the factor analyses results and high Spearman correlation >.8, this was duplicated under each theme. This document was circulated prior to the Scientific Group meeting. On the day, a preparatory session was conducted with the Expert Users Group and the clinicians separately to enable members to become familiar with all the materials to enhance their participation in the subsequent meeting. At the Scientific Group meeting, the items were considered with a view to selecting the best item or items.
The criteria for selecting items was to choose those that were qualitatively preferred and with the best psychometric properties. The principles for item selection were as follows:

a. Choose one item per theme to retain the face validity of the measure.

b. Decide if a second item is needed in this theme; and if so, choose a second item.

c. Retain a mix of negatively and positively worded items in the measure.

3. Results

3.1 Qualitative evidence

3.1.1 Qualitative evidence from adult service users

Items that fell under the following categories were excluded: items not relevant, items that were difficult for service users to respond to; ambiguous items (e.g. items whose meaning was not clear; had more than one meaning), distressing or sensitive items, and judgemental items (e.g. items that imply a certain belief or way of life). Full results of this stage are discussed elsewhere [13].

3.1.2 Qualitative evidence from young service users

The interviews revealed that while most of the ReQoL themes and items were seen as relevant to the younger service users, there was some indication of a slightly different perspective in relation to some of the themes and items. For example, none of the young people interviewed were living independently; the majority were living with birth parents. They described how living with older adults, who were to some degree responsible for their care, impacted on the degree of choice, control and autonomy they had. In relation to the self-perception theme, one participant reflected on a degree of confusion about oneself being a normal part of adolescent development. Two final other characteristics of the younger group were a tendency, when talking about their hopes for the future, to focus on academic achievements and motivation and, when discussing the physical health question, there was a tendency to equate physical health with physical fitness and healthy eating. None of the
young people had experienced prolonged physical health problems and so the question may have seemed less relevant to them.

3.1.3 Qualitative evidence from clinicians

The clinicians involved in the focus groups showed a preference for items that used words, phrases and concepts that they thought were frequently used in conversations with service users. They liked items that were realistic and related to people’s everyday lives, or picked up on issues that they considered important to the people they worked with. The clinicians sought to identify the items that related to functioning and recovery.

Some clinicians stated a preference for items that resonated with things they usually discussed, or those that could lead to further conversations. Many clinicians valued the inclusions of questions that could pick up risk and suicidal ideation, as this was something they actively monitored. Clinicians identified certain items from the item set that they did not ask about in their consultation but which they would find helpful as they may indicate the cause and or impact on the mental health components, for example the item ‘I had problems with my sleep’ and the physical health item.

There was debate about whether items should be general or specific, and also about items that required an assessment of other peoples' thoughts, feelings or intentions, such as “I felt people did not want to be around me”. Some questions were seen as having the potential to be skewed for particular group of service users. Items that presented a spectrum were preferred to those that presented absolutes such as “Everything in my life felt bad”. Some of the clinicians disliked words that they thought some people might not use or understand and those that had more than one meaning. Whenever there were conflicting views regarding the latter between the clinicians’ and the service users’ views, we adopted the views of the service users.

3.1.4 Results from the translatability assessment
In total, on sets of items in both waves (n = 98 items), 33 items were classified under a R1 recommendation (no change); 14 under R2 (no change, alternative wording for translation); 19 under R3 (change original), and 37 under R4 (delete item). Five items were put in two classifications (R3 and R4) as the linguists could not decide upon deletion or rewording. Reasons for suggesting deletion were either because of redundancy with other items, inadequacy with response categories, or the strong idiomatic nature of the original. When items were found to duplicate others, the suggestion for deletion was accompanied with a question mark to indicate that a decision had to be made as to which one to delete or keep.

Based on the results of the TA, below are some examples that affected final item selection:

- Rejection of the following idiomatic items ‘I had reasons to get out of bed in the morning’, ‘I found it hard to stand up for myself’, ‘I felt OK about myself’, ‘I could not bounce back from my problems’
- Rephrasing of the item ‘I felt at ease with who I am’ to ‘I felt at ease with myself’ because of a semantic issue
- Rejection of items with structural issues ‘I disliked myself’, ‘I felt unsure about myself’ as it was deemed that they might call for translations using a negative verbal form (such as ‘I did not like’, ‘I did not feel sure’).

(See supplementary materials for more details).

3.2 Psychometric evidence

When we analysed the item sets as completed by service users, none of the items had more than 5% missing data and therefore no item was dropped for that reason. The details of the psychometric studies are presented elsewhere [7, 15]. In summary, from the factor analysis, 12 highly correlated items highlighting potential redundancy were identified in Study 1 and in Study 2 (Table S4 supplementary materials). From the IRT analyses, in Study 2, four items were identified as misfitting in three of the four datasets. The latter items were: ‘I felt at ease with who I am’, ‘I could do the things
I wanted to do’, ‘I had the opportunity to do the things I wanted’ and ‘I felt safe’. None of the items exhibited any differential item functioning. In Study 2, two items were identified as being insensitive to change ‘I felt angry’ and ‘I thought people cared about me’. The information functions were generated and scrutinised by theme and the items ranked in order of how much information they provided, whether items were of a middling nature or provided information at extremes. The items providing the poorest information were: ‘I had the opportunity to do the things I wanted’, ‘I thought people cared about me’; ‘I felt angry’; ‘I felt hopeful about my future’ and; ‘I had problems with my sleep’.

3.3 Combining evidence

3.3.1 Study 1

In Study 1, members of the Scientific Group were presented with this information in advance by theme. An excerpt for the Self-perception theme can be found in Table 2. From the deliberative process, 22 items were deleted at that stage and one item ‘I felt miserable’ was added.
### Table 2 Combining the evidence – Study1 (Self-perception theme)

<table>
<thead>
<tr>
<th>IRT analyses</th>
<th>Qualitative evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spearman Correlation within theme &gt;0.7</strong></td>
<td><strong>Mean IRT value</strong></td>
</tr>
<tr>
<td>I felt unsure of myself</td>
<td>Not covered</td>
</tr>
<tr>
<td>I tended to blame myself for bad things that have happened</td>
<td>I felt like a failure</td>
</tr>
<tr>
<td>I felt like a failure</td>
<td>I disliked myself</td>
</tr>
<tr>
<td>I felt confident in myself</td>
<td>I am at ease with who I am</td>
</tr>
<tr>
<td>I felt at ease with who I am</td>
<td>I valued myself as a person</td>
</tr>
<tr>
<td>I valued myself as a person</td>
<td>I felt ok about myself</td>
</tr>
<tr>
<td>I disliked myself</td>
<td>I felt ok about myself</td>
</tr>
<tr>
<td>I felt confused about who I am</td>
<td></td>
</tr>
<tr>
<td>I felt ok about myself</td>
<td></td>
</tr>
</tbody>
</table>

Key: ✓ - Fine to select; ✗ - Not recommended for selection ? – mixed; F (for) A (against) M (mixed) – all figures represent percentages; Not covered – this item was not seen by this group.
3.3.2 Study 2

For Study 2, the qualitative and quantitative evidence were colour-coded as illustrated in Figure 2.

Figure 2: Combining qualitative with quantitative evidence (theme: hope subtheme: hopelessness)

In each theme, the best item based on the colour-coding was first chosen as shown in Table 3. Following a discussion, a decision was reached as to whether a second item under that theme was needed and if so, the second item was chosen using the same procedure. Two themes – self-perception and wellbeing – contain one item only. In the other themes where two items were deemed important, items were chosen such that one was positively worded and one negatively worded.
Table 3 Items chosen under each theme

<table>
<thead>
<tr>
<th>Theme</th>
<th>Item code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activity</strong></td>
<td>ACT1</td>
<td>I found it difficult to get started with everyday tasks</td>
</tr>
<tr>
<td></td>
<td>ACT2P</td>
<td>I enjoyed what I did*</td>
</tr>
<tr>
<td><strong>Belonging and relationships</strong></td>
<td>BEL2</td>
<td>I felt lonely</td>
</tr>
<tr>
<td></td>
<td>BEL3P</td>
<td>I felt able to trust others*</td>
</tr>
<tr>
<td><strong>Choice, control and autonomy</strong></td>
<td>CHO4</td>
<td>I felt unable to cope</td>
</tr>
<tr>
<td></td>
<td>CHO1P</td>
<td>I could do the things I wanted to do*</td>
</tr>
<tr>
<td><strong>Hope</strong></td>
<td>HOP4</td>
<td>I thought my life was not worth living</td>
</tr>
<tr>
<td></td>
<td>HOP1P</td>
<td>I felt hopeful about my future *</td>
</tr>
<tr>
<td><strong>Self-perception</strong></td>
<td>SEL2P</td>
<td>I felt confident in myself</td>
</tr>
<tr>
<td><strong>Wellbeing</strong></td>
<td>WB11</td>
<td>I felt happy</td>
</tr>
</tbody>
</table>

*Second item to have been chosen under each theme

In the **Choice** theme, the second item ‘I could do the things I wanted to do’, chosen in the Scientific Committee, had poorer (red) psychometric properties as the item was misfitting in addition to having an inferior information function. The item ‘I felt hopeful about my future’ was also chosen despite having the worst information function in that theme. Both items were discussed at length during the meeting and in the end, a consensus was reached. In the interest of content and face validity, we chose items that may not have been chosen if only the psychometrics evidence was being taken into consideration.

4. Discussion

This article set out to provide an account of how the concept of personal recovery can be conceptualised in a personal and socially meaningful way and to convey a method for capturing this view via the co-production of a measure of recovery of quality of life by service users and researchers using qualitative as well as quantitative approaches. Broadening the definition of recovery requires an equivalent broadening of the research methods used by combining qualitative and quantitative evidence when developing an outcome measure. We found that it was possible to present the multitude of often complex and technical evidence concisely so that service users, clinicians and
researchers with varying backgrounds in multi-disciplinary teams could equally contribute meaningfully to the final item selection.

The use of qualitative and quantitative methodologies is the hallmark of a mixed methods approach and has been widely adopted in the research literature [16]. However, it has been less well used in pursuit of measure development. Indeed, virtually all measure development depends solely on quantitative methods with the evidence for validity derived from psychometric analyses. Although we did not follow any formal model for combining qualitative and quantitative methods, our approach has close parallels to the framework advocated by Luyt [17] that, in turn, extended a model initially developed by Adcock and Collier [18]. Luyt’s framework identifies three different levels (theory, domains, and items) that are informed and then refined through an iterative process of combining qualitative and quantitative data. Our informal model broadly followed these levels (termed stages in our approach) with each stage informed by feedback from qualitative analyses (as well as quantitative data), which derived from service users. An interesting component of Luyt’s framework is that the central concept of validity is viewed as being established across methodologies (i.e., both qualitative and quantitative research) rather than multiple aspects of validity (e.g., construct, concurrent, discriminative) being determined within only quantitative methods.

It is obvious that that the qualitative evidence could not have been generated by the researchers alone. The co-production of the ReQoL clearly shows that it is imperative to include service users with lived experience in the development of a measure that is to be relevant to them. The process of excluding service users from the construction of measures is an extension of their exclusion from other activities, and thereby increasing their social isolation. By contrast, our view is that service users as well as clinicians, linguists and researchers all need to be included in the process of production as they all have their own perspective, life experiences, expertise and biases. Therefore, co-production is not necessarily the most straightforward way of constructing an outcome measure, but is by far the best
way to guarantee a more relevant one. As will have been clear, the ReQoL measures were co-produced in partnership by service users and researchers [6]. However, co-production is only a first step towards social inclusivity whereby service users challenge services to both adopt and implement measures for which service users have a sense of joint ownership. In this sense, the development of measurement tools becomes one other aspect within the area of mental health in which the views of service users need to be central.

While it is paramount to recognise the importance of face and content validity as enhanced by multiple perspectives, it is crucial that an outcome measure assesses what it purports to measure – that is, ensuring that the measure retains the necessary psychometric standards. As shown in Section 3.3.2, one misfitting item was selected based on qualitative evidence and the deliberation of the members at the Scientific Group meeting. Similarly, an item with a relatively poor information function was also chosen. It is therefore recognised that there has to be a trade-off between superior psychometric properties and the face and content validity. The key in constructing the best possible outcome measure lies in the ability to find the right balance between the two. We think that this has been achieved as the use of the ReQoL is increasing rapidly and the initial psychometric results are encouraging [7].

It has been recognised for a long time that qualitative and quantitative can inform each other and be very complementary [19]. However, in health a lack of integration of the two was recognised [20]. The value of this paper is that, instead of keeping the qualitative and the quantitative strands separately, it demonstrates in detail how the two have been successfully integrated. While it is not uncommon, for qualitative work to be carried out alongside psychometrics in measure development, to our knowledge, this is the first paper to use a diagrammatic approach and provide details on the process of integrating the two. The traffic light approach taken by Study 2 is a clear improvement on
the tabular depiction in Study 1 and was proved easier to understand by all the members of the Scientific Group.

There are various ways of using graphics [21] but in this paper we have shown a simple and effective way of presenting complex information to individuals with a view of empowering everyone to participate fully in decision-making. Our evidence shows that it is possible to combine the results from qualitative experts, analysts (psychometricians), clinicians and service users. However, it is important to be realistic as this approach involves greater planning, time and resources. This way of presenting information can be adopted in many areas and is reasonably generalisable.

One caveat regarding the study is that the various groups of adult service users, young service users, clinicians and linguists assessed different item sets since it was part of an iterative process. For example, at the end of the interviews with the service users, the items that were slightly reworded were not reassessed by the linguists. However, care was taken to ensure that the revised items were in line with the comments received about them.

5. Conclusions

The work reported in this article demonstrates that the co-production of the ReQoL outcome measures has been enhanced substantially by combining qualitative and quantitative evidence. It is reasonable to suggest that this method is applicable to outcome measures in general. While it is a valuable process, researchers need to be realistic about the resource implications and be aware of possible trade-offs. We have been able to produce two versions of the ReQoL measure that can be used to assess outcomes of people with mental health conditions and with future work the ReQoL will also be suitable for calculating quality adjusted life years in the conduct of economic evaluations.
Supplementary Materials: The following are available online at XXXX, more details on the translatability assessment; Tables S1: Results of Translatability Assessment of first set of items (wave 1 n = 87); Table S2. Results of Translatability Assessment of second set of items (wave 2 n = 11); Table S3. Pairs of possible redundant items identified in the factor analyses.

Acknowledgements: This is an independent research commissioned and funded by the Policy Research Programme in the Department of Health. The study was undertaken by the Policy Research Unit in Economic Evaluation of Health and Care Interventions (EEPRU) funded by the Department of Health Policy Research Programme. This independent research is 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). The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health.

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. 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. Special thanks to Andrew Grundy, Rob Hanlon, Jo Hemmingfield, John Kay, Dan Robotham and Diana Rose, for providing us with service users’ perspectives.

Author contributions: Anju Keetharuth drafted the manuscript and led the ReQoL project. Elizabeth Taylor Buck proposed the pictorial format of the evidence, conducted the interviews with young service users and clinicians and contributed to the writing of the relevant sections. Janice Connell conducted and analysed adult service interviews, and contributed to the writing of the relevant
sections. Katrin Conway and Catherine Acquadro conducted the translatability assessment and wrote the relevant sections in the manuscript. Jill Carlton conducted the interviews with service users. Thomas Ricketts conducted the interviews with the clinicians. Rosemary Barber chaired all the Expert User Group meetings. Michael Barkham advised on the design of the project and contributed significantly to the manuscript. John Brazier was the Chief Investigator of the project. All co-authors commented on the draft and approved the final draft.

Conflicts of interest: The authors declare no conflict of interest.

Ethics: Ethical approval was obtained from the Edgbaston NRES Committee, West Midlands (14/WM/1062). Informed consent was obtained from all participants in the study.

Abbreviations

The following abbreviations are used in this manuscript:

- **CFA**: Confirmatory Factor Analysis
- **DIF**: Differential Item Functioning
- **EFA**: Exploratory Factor Analysis
- **GP**: General Practice
- **GRM**: Graded Response Model
- **IRT**: Item Response Theory
- **NHS**: National Health Service
- **PROM**: Patient Reported Outcome Measure
- **QoL**: Quality of Life
- **ReQoL**: Recovering Quality of Life
- **TA**: Translatability assessment
REFERENCES


19. Archibald, M. M.; Radil, A. I.; Zhang, X.; Hanson, W. E., Current mixed methods practices in qualitative research: A content analysis of leading journals. *Int J Qual Methods* 2015, 14, (2), 5-33.
