Article:
Keetharuth, D. orcid.org/0000-0001-8889-6806, Rowen, D. orcid.org/0000-0003-3018-5109, Bjorner, J. et al. (1 more author) (2021) Estimating a preference-based index for mental health from the recovering quality of life measure: valuation of recovering quality of life utility index. Value in Health, 24 (2). pp. 281-290. ISSN 1098-3015

https://doi.org/10.1016/j.jval.2020.10.012

ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Preference-Based Assessment

Estimating a Preference-Based Index for Mental Health From the Recovering Quality of Life Measure: Valuation of Recovering Quality of Life Utility Index

Anju Devianee Keetharuth, PhD, Donna Rowen, PhD, Jakob Bue Bjorner, PhD, John Brazier, PhD

ABSTRACT

Background: There are increasing concerns about the appropriateness of generic preference-based measures to capture health benefits in the area of mental health.

Objectives: The aim of this study is to estimate preference weights for a new measure, Recovering Quality of Life (ReQoL-10), to better capture the benefits of mental healthcare.

Methods: Psychometric analyses of a larger sample of mental health service users (n = 4266) using confirmatory factor analyses and item response theory were used to derive a health state classification system and inform the selection of health states for utility assessment. A valuation survey with members of the UK public representative in terms of age, sex, and region was conducted using face-to-face interviewer administered time-trade-off with props. A series of regression models were fitted to the data and the best performing model selected for the scoring algorithm.

Results: The ReQoL-Utility Index (UI) classification system comprises 6 mental health items and 1 physical health item. Sixty-four health states were valued by 305 participants. The preferred model was a random effects model, with significant and consistent coefficients and best model fit. Estimated utilities modeled for all health states ranged from −0.195 (state worse than dead) to 1 (best possible state).

Conclusions: The development of the ReQoL-UI is based on a novel application of item response theory methods for generating the classification system and selecting health states for valuation. Conventional time-trade-off was used to elicit utility values that are modeled to enable the generation of QALYs for use in cost-utility analysis of mental health interventions.

Keywords: mental health, preference-based measure, QALYs, ReQoL-10, ReQoL-20.

VALUE HEALTH. 2021; 24(2):281–290

Introduction

Quality adjusted life years (QALYs), a composite measure combining quality of life and duration of life, are used in cost-effectiveness analyses. Typically, the quality of life component of the QALY is generated using an off-the-shelf generic or condition-specific preference-based measure. The most commonly used generic preference-based measure, EQ-5D, has a focus on physical health (PH) with only 1 of the 5 dimensions directly pertaining to mental health (MH). There is growing evidence that EQ-5D is not well suited for use in certain areas of MH,[1–4] raising the question as to whether another preference-based measure with a larger focus on MH that also includes physical health, would be more appropriate for use in cost-effectiveness analyses in those areas. Such a measure may have the advantage of performing better psychometrically, as it may be better able to detect changes in MH over time and differences across treatments. In addition, the measure would be more relevant for and acceptable to people for inclusion in data collection with MH problems.

The Recovering Quality of Life (ReQoL) measures, ReQoL-10 and ReQoL-20, were developed for people aged ≤16 years old experiencing a broad range of mental health difficulties from common mental health problems to more severe psychotic ones.[5] They are intended for use in routine practice with people experiencing mental health difficulties and can be used to evaluate interventions for this population. ReQoL-10 contains 10 MH items and ReQoL-20, 10 additional items. Both versions contain a PH item (see Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012). The theoretical framework underpinning the themes for the measures were established from a qualitative literature review and in-depth interviews.[1,6,7] Six MH themes (activity; belonging; choice, control and autonomy; hope; self-perception; and well-being) and a theme relating to PH were identified. Psychometric evidence generated through 2 studies

recruiting more than 6500 service users[8,9] was combined with the qualitative evidence[10] to select items for the final measures.[11]

The aim of this article is to present the derivation of ReQoL utility index (ReQoL-UI), a recovery-focused generic preference-based measure derived from ReQoL-10 and ReQoL-20. It presents the development of a novel approach because standard methods used to select health states for valuation rely on independence between dimensions, which is not present between the MH items in ReQoL-UI.

Methods

ReQoL-UI was constructed in 4 stages: (1) the derivation of the classification system of ReQoL-UI; (2) the choice of health states for valuation using item response theory (IRT); (3) the time trade-off valuation (TTO) survey used to elicit values for a selection of ReQoL-UI health states; and (4) the modeling of preference weights that can be used to generate utility values for all health states defined by the ReQoL-UI.

ReQoL Data for Use in Stages 1 and 2

Data were gathered from 4266 individuals accessing MH services from primary (27%) and secondary care (67%), from a trial cohort for a depression study (5%) and from the voluntary sector (1%). The sample is described in detail elsewhere.[5,9] In summary, 58% of the sample was female, the age range was 16 to 98 and mean (SD) age was 47 (17) years. Respondents self-reported a wide range of diagnoses including common MH disorders (51%) and psychotic disorders (18%).

Stage 1: Development of the ReQoL-UI Classification System

The aim of this stage is to generate a health state classification system amenable for valuation. The 10 ReQoL-10 MH items and the PH item were therefore considered for the reduced classification system as these items also appear in ReQoL-20. However, the use of all the 11 items to elicit preference weights during the valuation exercise would be cognitively too onerous.[12] To maintain the face validity of the ReQoL-10 measure, we chose 1 item from each of the 6 MH themes and the PH item, all being identified as important to services users experiencing MH difficulties. To select the MH items, we adopted the following steps: (1) consider the dimensionality of the ReQoL item set; (2) exclude any misfitting item(s); (3) select items with the best psychometric properties. For step (1) confirmatory factor analysis was undertaken using MPlus 7.3.[13] Model fit was assessed using root mean square error approximation and comparative fit index. In a bifactor model providing an adequate fit, the negatively (n = 24) and positively worded items (n = 15) loaded onto a "negative" factor and a "positive" factor, respectively.[9] However, the explained common variance of the global factor was 85% suggesting the measure could be appropriately analyzed using unidimensional IRT models.

To undertake steps (2) and (3), the graded response IRT model was fitted to the 39 items to estimate item parameters and the full results are presented elsewhere (see Keetharuth et al[8] for full results). The graded IRT model expresses the probability of a particular response to a ReQoL item as a function of item characteristics (item discrimination and item thresholds) and a latent mental health variable (theta [θ]), which is assumed to have a standard normal distribution (with high scores indicating good mental health). Based on the graded response model, θ can be estimated for each respondent and the contribution of each item to the overall measurement precision at a given θ level can be assessed through Fisher information functions.[14] For various levels of θ scores ranging from -2 to 2 in intervals of 0.4, the ReQoL-10 items were ranked in order of the item's contribution to measurement precision. This approach ensured that the most informative items were chosen and that the items covered the range of severity observed among MH service users. IRT analyses were carried out using IRTPRO 3.1.[15]

Stage 2: Selecting Health States

Standard approaches for selecting health states (eg, orthogonal arrays) for valuation, rely on independence between dimensions, which is not the case in ReQoL-UI. Previous studies where the classification system has a unidimensional component with correlated items have used a Rasch vignette approach.[16,17] The latter approach uses Rasch-based threshold analysis to select commonly observed health states for valuation, and then generates utility values for all possible health states using a regression model that predicts TTO utilities using the Rasch score for the health state. Here, we adapted this approach to IRT methods rather than Rasch analysis because IRT models have been shown to provide a good description of the ReQoL items,[8] and IRT provides more flexibility in modeling than Rasch analysis.

We selected health states for valuation choosing the response combinations that are most likely to be encountered in practice by estimating the probability of each possible combination of health states according to the graded response model. We performed such calculations across the entire range of estimated θ values from -2.18 (worst score on all 6 items) to 1.85 (best scores on the chosen 6 items). To achieve a reasonable trade-off between complexity and detail, we categorized this range into 15 score groups: score group 1 through 8 covered the range from -2.18 to 0, while score group 9 through 15 covered the range from 0 to 1.85. For each score group, the response combinations providing a score within this range were ranked according to their probability and the 3 most likely response combinations were chosen as health states for valuation (see Appendix Fig. 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012). For score group 15, only one response combination (555555) provided a score in this range, so this score group only contributed one health state (for a total of 14*3+1=43 health states). To ensure accurate utility assessment of poor MH states, we purposively oversampled response combinations providing a MH score below the average. For each of the 8 score groups below 0, we selected 2 additional response combinations for a total of 59 MH states (43+8*2). These were combined with the PH item by randomly selecting one physical level to be considered together with the MH states (using the random number generator in Excel). Five additional combinations of PH and MH states were added. For mental health, these included the worst possible MH score (555555), the best possible score (111111), and a score indicating "average" MH (333333). This approach was chosen because MH and PH form 2 separate dimensions and appear separately in the regression analyses undertaken in stage 4. All items were scaled from 1 to 5 with level 5 indicating the worst PH or MH (highest level of impact).

Stage 3: Design and Conduct of the Valuation Study

People's preferences for the sample of health states previously selected were elicited using TTO, a choice-based technique, in face-to-face interviews with members of the UK public. Based on similar valuation studies, we intended to recruit around 300 participants.[18] Respondents were selected to generate a nationally representative sample based on age and sex from postcodes in Scotland, England, and Wales. Households in the selected areas

received a letter in advance, advising them that an interviewer would call, with an opportunity to opt out. Interviews were managed by a market research agency and were conducted by experienced interviewers trained by the researchers. Face-to-face training was provided to all interviewers by 2 experienced researchers. First, an overview of the task was presented. Second, the interviewers were familiarized with the health state classification. Third, role-play in pairs provided each interviewer with an opportunity to interview someone else and be interviewed. Each interviewer was assigned a supervisor who could answer any questions and provide support; preliminary data checks were also carried out by the agency. Interviews were held in the respondents' own home and respondents were offered £10 for their participation.

During the interview, respondents first completed demographic and health questions followed by the ReQoL-10 to familiarize themselves with the health state classification system and response options. The interviewer told the participants that the health states were made up using statements from the questions you had just seen. Second, respondents undertook a warm-up task in the form of a practice TTO question. The interviewer had the discretion to decide whether a second practice question was necessary. Third, respondents undertook TTO valuation of 8 different health states. The Measurement and Valuation of Health protocol and its related props were used for states better than dead,[19] and lead-time TTO was used for states worse than dead.[20] This approach is the composite time trade-off approach that is in accordance with the protocol used internationally to value the EQ-5D-5L, that was developed to resolve the issue that previous TTO protocols required an arbitrary rescaling of states valued worse than dead.[21] Respondents were first asked whether they would prefer to live in the health state to be valued for 10 years and then die, or to die immediately to establish whether the health state was better, worse, or equal to being dead. For health states better than dead, participants were asked to imagine they would be in the health state that was being valued for a period of 10 years. They were then asked to consider a number of shorter periods in full health (x) to ascertain how many years of full health the respondent was willing to give up to avoid being in the impaired health state that was being valued. At the point where respondents were indifferent between x years in full health and 10 years in the state, the state took the value $x/10$. For states worse than dead, lead-time TTO was used which involves the same approach but adds a lead-time of 10 years to both full health and the impaired health state to allow respondents to trade these 10 years to avoid the impaired health state.[20] The state took the value $-y/10$, where y is the number of lead-time years that the respondent is prepared to sacrifice to avoid the impaired health state. Finally, respondents rated how difficult they found the tasks, and interviewers rated how well they thought the respondent had understood and engaged with the task.

Stage 4: Modeling Health State Preferences

The ReQoL-UI MH items form a unidimensional MH component, with the PH item constituting a second dimension. Therefore, similar to the modeling approach used in the Rasch vignette approach,[22] TTO values were regressed on the IRT-based MH score (estimated through the expected *a posteriori* approach) and dummy variables to represent 4 of the severity levels of the PH item (with level 1 as the reference case). The expected a posteriori estimates (θ) were rescaled from a range of -2.18 to 1.85 to a scale of 0 (best possible mental health) to 1 (worst possible mental health). Different regression models were fitted using mean and individual level data including a simple linear relationship, quadratic and cubic relationships. First, model specifications included mean level ordinary least squares (OLS) where mean scores were regressed on the rescaled θ scores and on dummy variables for the levels of the PH item. To account for multiple observations per individuals we also estimated random effects (RE) models[23] using maximum likelihood estimation. The error term $\varepsilon_{ij} = u_j + e_{ij}$ where u_j is the random effect and e_{ij} represents the random error term for the ith health state valuation of the jth individual.

Table 1. Evaluation of most informative item by each score level.

θ	Item information functions									Most informative item on each score level ranked by iteration				
	ACT1	ACT5P*	BEL2*	BEL3P	CHO4*	HOP1P	HOP4*	SEL2P*	WB11P*	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
−2	0.703	1.034	0.700	0.648	0.923	0.578	1.305	0.410	0.749	HOP4*	ACT5P*	CHO4	WB11P	ACT1
−1.6	0.953	1.571	1.143	0.731	2.319	0.868	2.109	0.976	1.576	CHO4*	HOP4	WB11P	ACT5P	BEL2*
−1.2	1.077	1.753	1.443	0.764	3.251	1.065	2.413	1.731	2.159	CHO4	HOP4	WB11P	ACT5P	SEL2P
−0.8	1.114	1.783	1.537	0.778	3.373	1.133	2.475	2.081	2.179	CHO4	HOP4	WB11P	SEL2P	ACT5P
−0.4	1.124	1.745	1.560	0.784	3.370	1.154	2.455	2.140	2.277	CHO4	HOP4	WB11P	SEL2P	ACT5P
0	1.124	1.735	1.562	0.785	3.402	1.157	2.145	2.158	2.231	CHO4	WB11P	SEL2P	HOP4	ACT5P
0.4	1.120	1.821	1.519	0.779	3.271	1.166	1.330	2.187	2.304	CHO4	WB11P	SEL2P	ACT5P	BEL2
0.8	1.098	1.800	1.323	0.754	2.283	1.160	0.581	2.203	2.284	WB11P*	CHO4	SEL2P	ACT5P	BEL2
1.2	1.027	1.534	0.913	0.685	0.897	1.084	0.211	1.968	2.184	WB11P	SEL2P*	ACT5P	HOP1P	ACT1
1.6	0.833	0.961	0.495	0.554	0.262	0.874	0.071	1.269	1.536	WB11P	SEL2P	ACT5P	HOP1P	ACT1
2	0.554	0.458	0.231	0.393	0.070	0.579	0.023	0.581	0.716	WB11P	SEL2P	HOP1P	ACT1	ACT5P

The following were not selected: ACT1 "I found it hard to get started with everyday task," BEL3P "I felt able to trust others," HOP1P "I felt hopeful about my future." The remaining tenth item "I could do the things I wanted to do" was a misfitting item.
*Most informative items chosen for the health state classification system: ACT5P "I enjoyed what I did," BEL2 "I felt lonely," CHO4 "I felt unable to cope," HOP4 "I thought my life was not worth living," SEL2P "I felt confident in myself," and WB11P "I felt happy."

$$y_{ij}^{s} = f(\theta, X_{\lambda}, \beta) + \varepsilon_{ij}^{s}, \quad y_{i,j}^{s} = \begin{cases} \left\{1 - \frac{z_{ij}}{w_{ij}}\right\} & \text{if state better than dead} \\ \left\{1 + \frac{z_{ij}}{10}\right\} & \text{if state worse than dead} \end{cases}$$

Where i = 1,2 … n represents the individual health states and j = 1,2 … m represents the respondents. The dependent variable y_{ij}^{s} is disutility (1-TTO) for health state i valued by respondent j and θ represents IRT scores for the corresponding health state, X is a vector of dummy explanatory variables for each level λ of the PH items with level $\lambda = 1$ acting as a baseline. All models excluded a constant because we used full health as defined by ReQoL-UI level 1111111 as our upper anchor for TTO.[24] We explored the inclusion of interaction terms that interacted the severity of the MH component, θ, with the PH dimension, where, as health worsens the interaction term increases. We estimated consistent models, where adjacent inconsistent levels of the physical dimension were merged, to ensure that as health worsened the utility value would not increase. All modeling was performed using STATA 15.[25]

Model Performance

Several criteria were used to evaluate model performance: (1) inconsistencies in parameter estimates and significance of coefficients; (2) comparing predictive model performance using root mean square error (RMSE), mean absolute error (MAE), difference between actual and predicted values at health state level, percentage of observations with absolute errors (AE) >0.05 and >0.1; and plots of actual and predicted health state values; (3) comparing Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for different model specifications within the same types of models.

Compliance With Ethical Standards

Ethical approval for stages 1 and 2 was granted by the Edgbaston National Research Ethics Services committee, West Midlands (14/WM/1062). Ethical approval for the valuation survey was obtained from the School of Health and Related Research via the University of Sheffield Research Ethics Committee process (reference number: 009307). Informed consent was obtained from all respondents in the study.

Results

Stage 1: Health State Classification System

Table 1 reports analyses of the ReQoL-10 items, using the IRT results. One item "I could do the things I wanted to do" was excluded as it was misfitting, indicating that standard IRT scoring may not be appropriate for this item. Through ranking the remaining 9 items in order of highest information across different severity levels for mental health, 5 items were first selected; items providing the most information at the highest severity levels were: "I thought my life was not worth living," and "I felt unable to cope," and "I enjoyed what I did"; the items providing the most information at the low severity end were: "I felt happy" and "I felt confident in myself." To ensure that each theme was represented, a sixth item from the belonging theme, "I felt lonely" was chosen even though it was the fifth best item at both the severe and milder ends. The selected items were rephrased to the present tense (Table 2).

Stage 2: Selecting Health States

The method described above yielded 59 health states for valuation (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012). The additional 5 health states provided the opportunity to add the pits state "5555555" as the best state was already selected. Two other states were selected so that more severe levels of physical impairment were combined with the best MH state to isolate the impact of PH more clearly; and finally, a moderate severity state was added.

Stage 3: Design and Conduct of the Valuation Study

Valuation survey participants

Based on similar studies, we recruited 305 participants.[18] Interviews were carried out by 15 experienced and trained interviewers each undertaking between 5 and 30 interviews. The proportion of total suitable participants answering their door at the time of the interview was 28%. Five participants were excluded from the analysis: 4 valued all health states as identical and less than 1, implying they did not understand the task; one valued all health states as worse than dead, implying that the participant thought that no state was worth living. The characteristics of the sample are compared with the population from England and Wales (Table 3).

Forty-eight respondents (16%) reported a MH condition, out of whom 35 were receiving treatment. The 3 most reported MH conditions were anxiety, depression, and stress-related (including posttraumatic stress disorder). One hundred respondents (34%) reported a physical problem with the 3 most reported conditions being high blood pressure, tiredness and fatigue, and pain. Only 5 (2%) and 29 (10%) respondents reported that they found the questions very difficult and quite difficult to understand, respectively. Interviewers noted that 5 (2%) respondents had not quite understood the questions; that 17 respondents (6%) did not concentrate very hard and had put little effort into the valuation task and that 2 respondents (<1%) concentrated at the beginning but subsequently lost concentration or interest. Interviews lasted 34 minutes on average (SD = 10).

Health State Values

The number of observations per state vary from 27 to 44. The distribution of observed TTO values show that 21%, 4%, and 6% of observations at 1, 0, and −1, respectively (see Appendix Fig. 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012). The mean observed TTO values by health state range from −0.178 (worst state = 5455555) to 0.966 (best state = 1111111). A measure-specific full health value below 1 is expected because the state is compared to "full health," which may be imagined by participants to be better than the state described. In the first 3 states (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012), it is evident that as PH severity increases, the mean TTO value falls. The worst health state (5555555) has higher mean TTO (−0.128) than the state 5455555 (−0.178) and state 5553554 (−0.144), but it should be noted that different respondents valued these health states.

Stage 4: Modeling the Health State Utility Data to Generate Utility Values for All Health States

The best performing mean linear and quadratic OLS and RE models, assessed in terms of MAE, RMSE, AIC, BIC, and observations with AE greater than 0.1 and 0.05 are presented in Table 4. The RE models were preferred to fixed effects models using the Hausman test. There were some inconsistent coefficients in the

Table 2. ReQoL descriptive system.

Theme	Description of health states	Levels
1. Activity (act5p: I enjoyed what I did)	I enjoy what I do most or all of the time	1
	I often enjoy what I do	2
	I sometimes enjoy what I do	3
	I only occasionally enjoy what I do	4
	I never enjoy what I do	5
2. Belonging and relationships (bel2: I felt lonely)	I never feel lonely	1
	I only occasionally feel lonely	2
	I sometimes feel lonely	3
	I often feel lonely	4
	I feel lonely most or all of the time	5
3. Choice, control and autonomy (cho4: I felt unable to cope)	I never feel unable to cope	1
	I only occasionally feel unable to cope	2
	I sometimes feel unable to cope	3
	I often feel unable to cope	4
	I feel unable to cope most or all of the time	5
4. Hope (hop4: I thought my life was not worth living)	I never think that my life is not worth living	1
	I only occasionally think that my life is not worth living	2
	I sometimes think my life is not worth living	3
	I often think my life is not worth living	4
	Most or all of the time I think my life is not worth living	5
5. Self-perception (sel2p: I felt confident in myself)	I feel confident in myself most or all of the time	1
	I often feel confident in myself	2
	I sometimes feel confident in myself	3
	I only occasionally feel confident in myself	4
	I never feel confident in myself none of the time	5
6. Wellbeing (wb11p: I felt happy)	I feel happy most or all of the time	1
	I often feel happy	2
	I sometimes feel happy	3
	I only occasionally feel happy	4
	I never feel happy	5
7. Physical health item (please describe your physical health: problems with pain, mobility, difficulties caring for yourself, or feeling physically unwell)	I have no problems with physical health	1
	I have slight problems with physical health	2
	I have moderate problems with physical health	3
	I have severe problems with physical health	4
	I have very severe problems with physical health	5

linear models for levels 2 and 3 of PH compared with level one. The coefficients for the quadratic models were all in the direction expected, where increasing severity leads to decreases in utility, with the exception of the interaction terms combining level 2 of PH and θ (compared with level 1) for both OLS and RE model 2, where the coefficient was positive rather than negative (for linear and quadratic RE complete model results, see Appendix Tables 2 and 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012). The cubic models are not presented as they do not provide a monotonous decreasing utility scores for worse MH (see Appendix Tables 4 and 5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012). We present a summary comparison of the models in Appendix Table 6 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012. The best performing mean level OLS model (model 2) and RE models (model 6) consist of a quadratic specification of θ with interaction terms for θ and levels 3, 4, and 5 of PH. They have the lowest RMSE, lowest AIC and BIC, and lowest percentage of observations with AE $<$0.1 and 0.05. The interaction terms in all models are negative. As shown in Figure 1 and Appendix Figure 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.10.012, neither models exhibit systematic bias in the predictions by severity for the majority of health states except for the most severe states where larger prediction errors were observed.

Table 3. Characteristics of respondents in the valuation survey.

	Mean	SD	Range	England and Wales norms
Age	51.6	19.1	18-96	39*
Life satisfaction score	8.0	1.8	2-10	7.5†
Health satisfaction score	7.7	2.0	1-10	

		n	Percentage (%)	England and Wales norms‡ (%)
Sex	Male	135	45.0	49.1
	Female	164	54.7	50.9
	Other	1	0.3	
Marital Status	Single	67	22.3	34.6
	Married/partner	161	53.7	46.6
	Separated/divorced	26	8.7	11.6
	Widowed	45	15.0	7.0
	Prefer not to say	1	0.3	
Ethnicity	White	278	92.7	86.0
	Asian/Asian British	16	5.3	7.5
	Black/African/Caribbean/Black British	3	1.0	3.3
	Other ethnic group	3	1.0	3.2
Degree	Yes	87	29.0	27.0
	No	202	67.3	
	Missing	11	3.7	
Main activity	Employed	146	48.7	61.7
	Retired	97	32.3	13.9
	Housework	22	7.3	4.3
	Student	5	1.7	9.3
	Unemployed	16	5.3	4.4
	Long-term sick	8	2.7	4.3
	Other	6	2.0	2.2
Overall health	Excellent	34	11.3	
	Very good	126	42.0	
	Good	96	32.0	
	Fair	31	11.4	
	Poor	10	3.3	
	Missing	1	0.3	
Age categories, y	16-25	20	6.7	11.9%
	26-64	173	57.7	52.8%
	≥65	84	28.0	16.4%
	Missing	23	7.7	
Experienced serious illness yourself	Yes	83	27.7	
	No	211	70.3	
	Missing	6	2.0	
Experienced serious illness in the family	Yes	143	47.7	
	No	149	49.7	
	Missing	8	2.7	
Experienced serious illness in caring for others	Yes	77	25.7	
	No	215	71.7	
	Missing	8	2.7	
How well interviewer thought the respondent understood and carried out the TTO tasks during the interview? (answered by interviewers)	Understood and performed exercises easily	192	63.37	
	Some problems but seemed to understand the exercises in the end	106	34.98	
	Doubtful whether the respondent understood the exercises	5	1.65	
Level of concentration and effort of the respondent as perceived by the interviewer	Concentrated very hard and put in a great deal of effort	143	47.19	
	Concentrated fairly hard and put in some effort	140	46.20	
	Did not concentrate very hard and put in little effort into it	18	5.94	
	Concentrated at the beginning but lost interest/concentration toward the end	2	0.66	

*Median age only was found.
†Office of National Statistics life satisfaction 2016.
‡Statistics for England in the Census 2011. The census includes persons aged ≥16 years, whereas this study only surveys persons aged ≥18 years.

RE model 6 is the overall preferred model because it had better predictive ability, albeit only slightly better than the OLS model when comparing the lowest proportion of absolute errors greater than 0.05 and 0.1. The estimates for the best health state and worst states are 1 and −0.195, respectively. Depicting the mean TTO predicted by model 6 for levels 2 to 5 of PH indicate that the decrements for the first 2 levels of the PH item are very similar, with by far the largest gap being between levels 3 and 4 of physical functioning (Fig. 2).

Discussion

We developed the ReQoL-UI health classification, which comprises 6 MH and 1 PH item from ReQoL-10 (and ReQoL-20) and have produced a set of preference weights. An algorithm has been estimated to generate the ReQoL-UI scores and available in STATA, SPSS, and Excel, using the predictions from the preferred RE model with the corresponding θ for all the possible combinations for the 7 items. The preference weights enable utility values to be generated from the ReQoL measures for use in cost-effectiveness analyses across the full severity range of MH conditions. ReQoL-10 and ReQoL-20 were specifically developed with considerable inputs from service users and have high face and content validity.[10,11,26] Therefore, the corresponding utilities are likely to be more appropriate for use to evaluate mental healthcare interventions than those generated from generic preference-based measures with a larger focus on PH rather than on MH.

Table 4. Regression results for estimating health preference scores.

	OLS mean models			Random effects models			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 (preferred)	Model 7
	Linear with interactions	Quadratic with interactions	Quadratic with interactions, only significant coefficients	Linear with interactions	Quadratic with interactions	Quadratic with interactions	Quadratic with interactions, only significant coefficients
θ (newtheta)	−0.433*	0.01	−0.053	−0.441*	0.028	0.028	−0.015
θ2 (newthetasq)		−0.572*	−0.517*		−0.582*	−0.581*	−0.558*
Phy2	0.059	−0.069		0.089	−0.033	−0.032	
Phy3	0.001	−0.073	−0.084*	0.027	−0.050	−0.049	−0.076*
Phy4	−0.140†	−0.284*	−0.270*	−0.141*	−0.265*	−0.265*	−0.261*
Phy5	−0.189*	−0.294*	−0.284*	−0.201*	−0.292*	−0.292*	−0.288*
Inter2	−0.099	0.066		−0.151	0.002		
Inter3	−0.135	−0.037		−0.165	−0.067	−0.067	
Inter4	−0.503*	−0.292*	−0.293*	−0.492*	−0.310*	−0.310*	−0.292*
Inter5	−0.501*	−0.362*	−0.356*	−0.465*	−0.350*	−0.351*	−0.330*
Constant‡	1	1	1	1	1	1	1
Observations	64	64	64	2303	2303	2303	2303
Adjusted R-squared	0.974	0.982	0.982				
RMSE	0.082	0.067	0.069	0.082	0.069	0.069	0.070
MAE	0.069	0.056	0.058	0.069	0.057	0.057	0.057
AIC	−121	−144	−147	3451	3430	3428	3426
BIC	−102	−122	−132	3514	3499	3492	3477
No. of observations with AE >0.1	15	9	8	13	8	8	10
Percentage of observations with AE >0.1	23%	14%	13%	20%	13%	13%	16%
No. of observations with AE >0.05	42	32	33	39	29	29	32
Percentage of observations with AE >0.05	66%	50%	52%	61%	45%	45%	50%

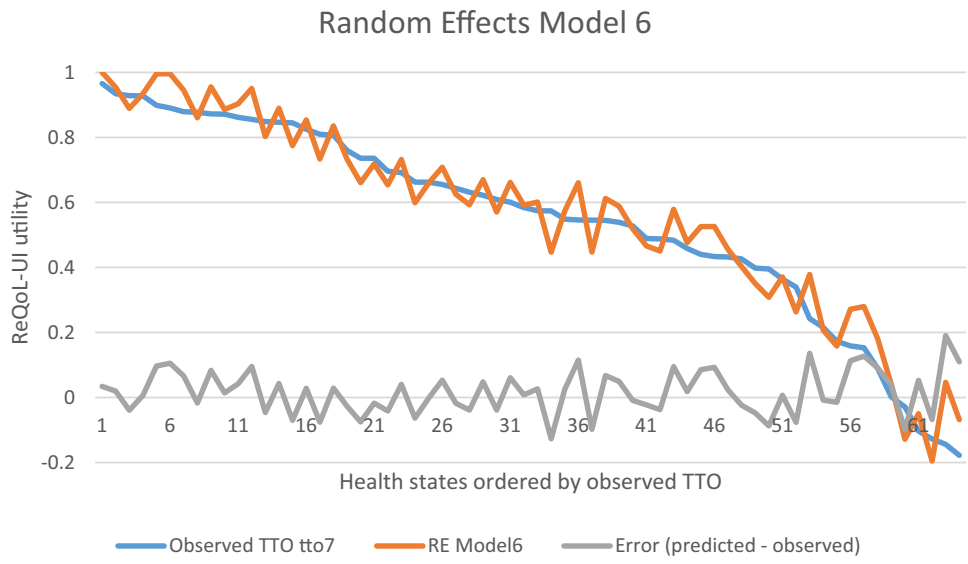
θ indicates IRT theta rescaled to 0 (best possible mental health score) and 1 (worst possible mental health scores); AE, absolute error; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; inter4 = θ * phy4 inter5 = θ * phy5; MAE, mean absolute error; phy2, level 2 physical health (phy1, best physical health and phy5, worst); phy3, level 3 physical health; phy4, level 4 physical health; phy5, level 5 physical health; RMSE, root mean square error.

*$P<.01$.

†$P<.05$.

‡The models did not have a constant but a constant 1 is presented here so the coefficients can be presented as utility decrement.

Figure 1. Plot of predicted versus observed utility values for the random effects (RE) model 6.

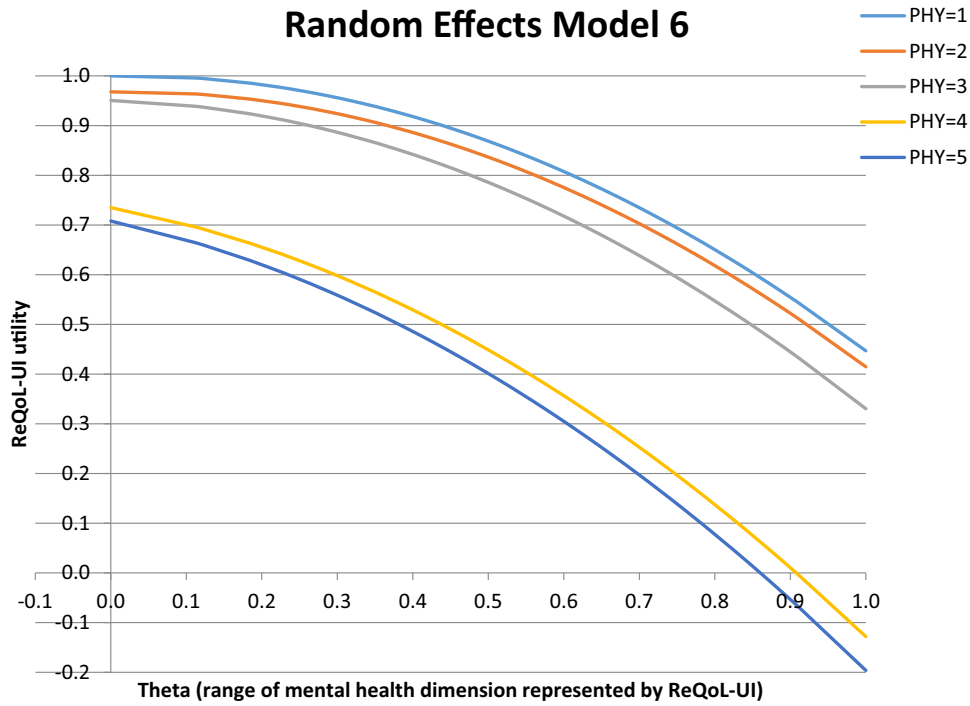


ReQoL-UI indicates Recovering Quality of Life Utility Index; TTO, trade-off valuation.

Although MAE is higher than some other TTO valuation studies where error is in the region of 0.05,[27,28] it is possible that this is due to the different model specification estimated here that does not differentiate between the different MH items in terms of their differential impact on utility. Both the mean OLS models and the RE models have good predictive ability across the range of utility values, with predictive performance lowest for all models for the very severe states. The poorer predictive performance for the models for the more severe states may have been observed due to the inconsistencies in the TTO utility values for some of the more severe states, where worst state had a similar but slightly higher mean TTO value than 2 other severe health states. Because the value set that generates utility values for all health states is based on modeled values, all utility values are logically consistent where utility either remains the same or is lower as the health state becomes more severe.

Unlike the EQ-5D, 6 of the items form a unidimensional component in the ReQoL-UI classification system related to MH,

Figure 2. Mean predicted trade-off valuation and for each level of physical health.



ReQoL-UI indicates Recovering Quality of Life Utility Index.

with one dimension for PH. From the regression results, the importance attributed to both PH and MH is clear. In the preferred RE model, more than 50% of the utility decrement is attributed to the severity of the MH condition compared with 23% for the worst level (level 3) of anxiety and depression in the EQ-5Q-3L preference weights.[19] There is a possibility that this may be in part either due to a framing effect or due to the number of MH items in the classification system proportional to the number of PH items. However, previous research has shown that all aspects of health in a classification system do not always receive a sizable utility decrement (see for example 'appetite' in EORTC QLU-C10D UK weights[29] or 'worried' in CHU9D UK weights),[30] meaning that the presence of an item alone does not guarantee a utility decrement, nor does it follow that the larger the number of dimensions the lower the utility values for the health state. Nonetheless, the PH item has a large utility decrement of 0.29 for the most severe level. One key advantage of including the PH item is that a utility decrement is generated for PH as well as for mental health, and as our qualitative research showed PH should not be ignored for people with MH problems.[7,10] In all the models, the signs of the interaction terms are negative and highly significant. This finding means that association between poor MH (θ) and low utility values is stronger when there is a moderate to severe PH problem.

This article provides an innovative use of IRT to select items and health states for a preference-based classification system. It improves the credibility of the states selected for valuation compared to the use of statistical designs such as an orthogonal array that can generate states with unlikely combinations of levels across dimensions. Several articles have used item threshold based on Rasch analysis to construct such a health classification system,[16,17] but, to our knowledge, none have used analyses of response combination probabilities. This approach allowed us to choose the health states that are more likely to be observed in real life. We selected 59 MH states and this constitutes a clear advantage since the larger number of health states included in this valuation study provides for a more robust regression model compared with previous applications of this approach. We analyzed response combination based on the graded response IRT model, but the approach could also be applied with the Rasch model. Although the models are very similar, IRT models may fit a broader set of scales.

There are a number of potential concerns to the article. First, we only recruited 305 participants in the valuation survey. Although the sample size for TTO valuation studies carried out online tend to be much larger, several studies have similar or less participants.[18,31] The number of observations per state ranged from 27 to 44, which is lower than 100 recommended.[32] With 64 health states valued and each participant valuing 8 health states, interviewing 800 participants face-to-face, would have been prohibitive in terms of time and costs. A second set of concerns surround the spike in the TTO data at 0, 1, and -1. The spike at 1 is due to the classification system where some people may not be prepared to trade life years for at least some of the health states. The spike at 0 reflects that people would rather die than be in the impaired ReQoL-UI state, but are not prepared to sacrifice prior years of full health to avoid any time in the impaired health state. The spike at -1 is however caused by the TTO task because this is the lowest utility that respondents can provide in the task, and hence it may have been that for some respondents for some health states they would have expressed a lower utility value if they had been able.

Another concern raised with valuation of attributes that may be more condition-specific or symptomatic is possible focusing effects, where respondents can exaggerate their impact on utility as these have not been placed within the context of other symptoms or more generic aspects of health.[33,34] However, in this study the attributes are not condition-specific but rather focused on MH, and furthermore respondents considered PH problems alongside MH problems, which arguably may have minimized focusing effects on MH.

The ReQoL-UI can be used in cost effectiveness analyses to capture the utility impact of problems in MH. The choice of preference-based measures to inform policy is one that is debated, as many reimbursement agencies recommend the use of a generic preference-based measure, often citing a recommended measure.[34] For example, in the UK, the National Institute for Health and Care Excellence recommends the use of one particular measure, the EQ-5D for use in cost-effectiveness analyses for health technology assessment.[35] However, alternative preference-based measures can be used in sensitivity analyses and where it can be evidenced that EQ-5D is not valid for the condition or patient population of interest. ReQoL-UI has the advantage, compared with other generic preference-based measures for use in people with MH problems, that it was developed with considerable input from MH service users and has 6 items capturing mental health. Generic preference-based measures such as EQ-5D focus on PH while including MH, whereas ReQoL-UI is arguably a generic preference-based measure that focuses upon MH, while including PH. Although this can provide an advantage for the evaluation of MH interventions, the introduction of a measure, such as ReQoL-UI with a different focus can cause issues of comparability across evaluations undertaken in PH and MH, particularly if EQ-5D or another generic preference-based measure is used for PH and ReQoL-UI for MH. However, comparability in evaluations across interventions can be maintained if EQ-5D is used in base case analyses, and ReQoL-UI or other measures are used in sensitivity analyses. Future research will empirically test the use of the ReQoL-UI in trials and studies, including comparison with preference-based generic measures including EQ-5D and SF-6D to compare their relative psychometric performance, and also explore the suitability of mapping to enable easier comparisons of evaluations conducted using the different measures.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2020.10.012.

Article and Author Information

Accepted for Publication: October 19, 2020

Published Online: November 27, 2020

doi: https://doi.org/10.1016/j.jval.2020.10.012

Author Affiliations: School of Health of Related Research, University of Sheffield, Sheffield, UK (Keetharuth, Rowen, Brazier); Optum Patient Insights, Johnston, RI, USA (Bjorner); University of Copenhagen, Copenhagen, Denmark (Bjorner).

Correspondence: Anju Devianee Keetharuth, PhD, School of Health of Related Research, University of Sheffield, 208 West Court, 2 Mappin Street, Sheffield S1 4DT, United Kingdom. Email: d.keetharuth@sheffield.ac.uk

Author Contributions: *Concept and design*: Keetharuth, Rowen
Acquisition of data: Keetharuth, Rowen
Analysis and interpretation of data: Keetharuth, Rowen, Bjorner, Brazier
Drafting of the manuscript: Keetharuth, Rowen, Brazier
Critical revision of the paper for important intellectual content: Keetharuth, Rowen, Bjorner, Brazier
Statistical analysis: Keetharuth, Bjorner
Provision of study materials or patients: Keetharuth
Obtaining funding: Keetharuth

Conflict of Interest Disclosures: Drs Keetharuth and Rowen reported receiving grants from the National Institute for Health Research during the conduct of this study. Dr Bjorner reported receiving personal fees from Optum Patient Insights outside the submitted work. No other disclosures were reported.

Funding/Support: This research was funded by the National Institute for Health Research Policy Research Program (Ref: PRP 104/0001). The views expressed are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgment: The authors would like to thank all the participants in the project, the staff who have been involved in the recruitment of participants, and all the members of the governance groups. We would also like to thank Donna Davis and Liz Metham for project management.

REFERENCES

1. Brazier J, Connell J, Papaioannou D, 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 Technol Assess*. 2014;18:vii-viii, xiii-xxv, 1-188.
2. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. *Value Health*. 2011;14:907–920.
3. Saarni SI, Viertiö S, Perälä J, et al. Quality of life of people with schizophrenia, bipolar disorder and other psychotic disorders. *Br J Psychiatry*. 2010;197:386–394.
4. Mulhern B, Mukuria C, Barkham M, et al. Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D. *Br J Psychiatry*. 2014;205:236–243.
5. Keetharuth AD, Brazier J, Connell J, et al. Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. *Br J Psychiatry*. 2018;212:42–49.
6. Connell J, Brazier J, O'Cathain A, et al. Quality of life of people with mental health problems: a synthesis of qualitative research. *Health Qual Life Outcomes*. 2012;10:138.
7. 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.
8. Keetharuth AD, Bjorner JB, Barkham M, Browne J, Croudace T, Brazier J. An item response theory analysis of an item pool for the Recovering Quality of Life (ReQoL) measure. Qual Life Res. Online ahead of print. https://doi.org/10.1007/s11136-020-02622-2
9. Keetharuth AD, Bjorner JB, Barkham M, et al. Exploring the item sets of the Recovering Quality of Life (ReQoL) measures using factor analysis. *Qual Life Res*. 2019;28:1005–1015.
10. Connell J, Carlton J, Grundy A, et al. The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Qual Life Res*. 2018;27:1893–1902.
11. Keetharuth AD, Taylor Buck E, Conway K, et al. Integrating qualitative and quantitative data in the development of outcome measures: the case of the Recovering Quality of Life (ReQoL) measures in mental health populations. *Int J Environ Res Public Health*. 2018;15:1342.
12. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol Rev*. 1956;63:81.
13. Muthen LK, Muthen BO. *Mplus User's Guide*. Sixth ed. Los Angeles, CA: Muthen & Muthen; 1998-2011.
14. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969;34(Pt. 2):100.
15. Cai L, Du Toit S, Thissen D. *IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling [computer software]*. Chicago, IL: Scientific Software International; 2017.
16. Mavranezouli I, Brazier JE, Young TA, Barkham M. Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems (CORE-OM). *Qual Life Res*. 2011;20:321–333.
17. Young TA, Rowen D, Norquist J, et al. Developing preference-based health measures: using Rasch analysis to generate health state values. *Qual Life Res*. 2010;19:907–917.
18. Rowen D, Rivero-Arias O, Devlin N, et al. Review of valuation methods of preference-based measures of health for economic evaluation in child and adolescent populations: where are we now and where are we going? *Pharmacoeconomics*. 2020:1–16.
19. Dolan P. Modeling valuations for EuroQol health states. *Medical Care*. 1997:1095–1108.
20. Devlin NJ, Tsuchiya A, Buckingham K, et al. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time'-approach. *Health Econ*. 2011;20:348–361.
21. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014;17:445–453.
22. Mavranezouli I, Brazier JE, Rowen D, et al. Estimating a preference-based index from the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) valuation of CORE-6D. *Med Decis Making*. 2013;33:381–395.
23. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002;21:271–292.
24. Yang Y, Brazier JE, Tsuchiya A, et al. Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. *Medical Decision Making*. 2011;31:281–291.
25. StataCorp S. *Statistical software: release 14*. College Station: StataCorp LP; 2015.
26. Grundy A, Keetharuth D, Barber R, et al. Public involvement in health outcomes research: lessons learnt from the development of the Recovering Quality of Life (ReQoL) measures. *Health Qual Life Outcomes*. 2019;17.
27. Rowen D, Brazier J, Young T, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value Health*. 2011;14:721–731.
28. Rowen D, Mulhern B, Banerjee S, et al. Estimating preference-based single index measures for dementia using DEMQOL and DEMQOL-Proxy. *Value Health*. 2012;15:346–356.
29. Norman R, Mercieca-Bebber R, Rowen D, et al. UK utility weights for the EORTC QLU-C10D. *Health Econ*. 2019;28:1385–1401.
30. Stevens K. Valuation of the child health utility 9D index. *Pharmacoeconomics*. 2012;30:729–747.
31. Mulhern B, Labeit A, Rowen D, et al. Developing preference-based measures for diabetes: DHP-3D and DHP-5D. *Diabet Med*. 2017;34:1264–1275.
32. Attema AE, Edelaar-Peeters Y, Versteegh MM, et al. Time trade-off: one methodology, different methods. *Eur J Health Econ*. 2013;14(Suppl 1):S53–S64.
33. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability? *Health Economics*. 2010;19:125–129.
34. Rowen D, Zouraq IA, Chevrou-Severac H, et al. International regulations and recommendations for utility data for health technology assessment. *Pharmacoeconomics*. 2017;35:11–19.
35. National Institute for Clinical Excellence (NICE). *Guide to the Methods of Technology Appraisal 2013*. London: National Institute for Health and Care Excellence; 2013.