Full title: Estimating cost-effectiveness using alternative preference-based scores and within-trial methods: exploring the dynamics of the QALY using the EQ-5D-5L and ReQoL-UI

Short title: The dynamics of the QALY; DOI: https://doi.org/10.1016/j.jval.2021.11.1358

Matthew Franklin (MF)[1], ORCID ID: 0000-0002-2774-9439
Rachael Maree Hunter (RH)[2], ORCID ID: 0000-0002-7447-8934
Angel Enrique (AE)[3,4], ORCID ID: 0000-0003-0585-4008
Jorge Palacios (JP)[3,4], ORCID ID: 0000-0002-2103-5507
Derek Richards (DR)[3,4], ORCID ID: 0000-0003-0871-4078

[1] Health Economics and Decision Science (HEDS), ScHARR, University of Sheffield, West Court, 1 Mappin Street, Sheffield, UK. S1 4DT.
[2] Research Department of Primary Care and Population Health, Royal Free Medical School, University College London, Royal Free Campus, Rowland Hill Street, London, UK.
[3] Clinical Research & Innovation, SilverCloud Health, Dublin, Ireland, UK.
[4] E-mental Health Research Group, School of Psychology, University of Dublin, Trinity College, Dublin, Ireland, UK.

Corresponding author
Dr Matthew Franklin, BA, MSc, PhD
Health Economics and Decision Science (HEDS), School of Health and Related Research (ScHARR), University of Sheffield, West Court, 1 Mappin Street, Sheffield. UK. S1 4DT.
Email: matt.franklin@sheffield.ac.uk; Tel: (+44) 114 222 4226

Summary: exploration of QALY and subsequent cost-effectiveness estimates based on different preference-based scores, mathematical and statistical methods as part of a case study within-trial economic evaluation.

Author Contributions: *Concept and design*: Franklin. *Acquisition of data*: Enrique, Richards. *Analysis and interpretation of data*: Franklin, Hunter, Enrique, Palacios, Richards. *Drafting of the manuscript*: Franklin, Hunter, Enrique, Palacios, Richards. *Critical revision of the paper for important intellectual content*: Franklin, Hunter, Enrique, Palacios, Richards. *Statistical analysis*: Franklin. *Obtaining funding*: Franklin, Enrique, Palacios, Richards.

Conflict of Interest Disclosures: Dr. Franklin reports other from SilverCloud Health, other from NIHR ARC YH, during the conduct of the study. Ms Hunter has nothing to disclose. Dr. Enrique, Dr. Palacios, and Dr. Richards are employees of SilverCloud Health. No other disclosures were reported.

Funding/Support: The trial from which the data for analysis was obtained, was funded by SilverCloud Health. Study resources for the trial from Berkshire NHS Foundation Trust, including R&D support, psychological wellbeing practitioners, case managers, and lead clinicians have been generously given in kind for the purpose of trial execution. The writing of the manuscript was part-funded by the NIHR Applied Research Collaboration Yorkshire and Humber (NIHR ARC YH; NIHR award identifier: 200166).

Role of the Funder/Sponsor: Employees of SilverCloud Health had a role in the design and conduct of the study; collection, management, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The NIHR 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; and decision to submit the manuscript for publication. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The funding agreement ensured the authors' independence in developing the purview of the manuscript, writing, and publishing the manuscript.

Acknowledgment: We would like to thank the R&D and clinical team members at Berkshire NHS Foundation Trust service for assisting trial execution: Judith Chapman, Gabriella Clark, Emma Cole, Sarah Sollese. We thank our colleagues at SilverCloud for providing administrative support and assisting data collection and analysis. We thank Anju Keetharuth, Donna Rowen, and John Brazier at ScHARR, University of Sheffield, for answering our questions in regards to the ReQoL-UI. We also thank the many patients who volunteered their time and efforts to participate in the trial.

Abstract

Objectives: to explore QALY and subsequent cost-effectiveness estimates based on the more physical health focussed EQ-5D-5L value set for England (VSE) or cross-walked EQ-5D-3L UK value set scores, or more mental health recovery-focussed ReQoL-UI, when using alternative within-trial statistical methods. We describe possible reasons for the different QALY estimates based on the interaction between item scores, health state profiles, preference-based scores, mathematical and statistical methods chosen.

Methods: QALYs are calculated over 8-weeks from a case study 2:1 (intervention: control) randomised controlled trial in patients with anxiety and/or depression. Complete-case (CC) and with missing cases imputed using multiple-imputation (MI) analyses are conducted, using unadjusted and regression baseline-adjusted QALYs. Cost-effectiveness is judged using incremental cost-effectiveness ratios and acceptability curves. We use previously established psychometric results to reflect on estimated QALYs.

Results: 361 (241: 120) people were randomised. EQ-5D-5L cross-walk produced higher incremental QALYs than VSE or ReQoL-UI which produced similar unadjusted QALYs, but contrasting baseline-adjusted QALYs. Probability of cost-effectiveness < £30,000 per QALY ranged from 6% (CC ReQoL-UI baseline-adjusted QALYs) to 64.3% (MI EQ-5D-5L cross-walk unadjusted QALYs). The control-arm improved more on average than the intervention-arm on the ReQoL-UI, a result not mirrored on the EQ-5D-5L nor condition-specific (PHQ-9, depression; GAD-7, anxiety) measures.

Conclusion: ReQoL-UI produced contradictory cost-effectiveness results relative to the EQ-5D-5L. The EQ-5D-5L's better responsiveness, and 'anxiety/depression' and 'usual activities' items drove the incremental QALY results. The ReQoL-UI's single physical health item and 'personal recovery' construct may have influenced its lower 8-week incremental QALY estimates in this patient sample.

Highlights

What is already known about the topic? The EQ-5D-5L has been questioned for the purpose of economic evaluation in mental health populations due to its more physical health focus, with a suggested need for a more mental health focussed measure to elicit QALYs. However, different preference-based measures and scores alongside mathematical and statistical methods for within-trial economic evaluation will produce different QALYs despite the "QALY is a QALY" assumption, impacting the subsequent cost-effectiveness results.

What does the paper add to existing knowledge? The ReQoL-UI is a relatively newer preference-based generic measure which focusses more on recovery-focused mental health than physical health as an alternative to the EQ-5D-5L for estimating QALYs in mental health service users. This is the first analysis to compare the EQ-5D-5L against the ReQoL-UI for estimating QALYs in a case study trial.

What insights does the paper provide for informing health care-related decision making? We explored the interaction between item scores, health state profiles, preference-based scores, mathematical and statistical methods, each of which played a part in the QALYs estimated. The aforementioned need to be considered holistically to understand and explain the different cost-effectiveness estimates as on QALY face value, these aspects are not transparent and so could mean inappropriate evidence may be used to inform decision makers when QALYs are considered equivalent and fully comparable.

Key Words: QALY, EQ-5D-5L, ReQoL-UI, cross-walk, anxiety, depression, recovery, economic evaluation, statistical methods.

1. Introduction

Economic evaluation evidence helps inform resource-allocation between alternative care interventions within a finite care budget (1). Cost-utility analysis (CUA) via cost per quality-adjusted life year (QALY) is recommended internationally, including by the National Institute for Health and Care Excellence (NICE) for England and Wales (2-4). QALYs are measured on a preference-based quality-adjustment scale, anchored at 0 (*a state equivalent to dead*) and 1 (*full health*), combined with length of life allowing comparisons between interventions that affect quantity and/or quality of life (1, 5). However, the concept of "a QALY is a QALY" for cross-comparable decision-making has been debated extensively given different preference-based measures and value sets produce different QALYs, stemming from aspects such as content and size of classification systems, and methods and populations used to value health states (5-12). Additionally, alternative mathematical and statistical methods can influence QALY estimates and associated cost-effectiveness evidence (13-15).

A more consistent, comparable approach is a rationale for NICE and reimbursement agencies internationally recommending the EQ-5D three-level version (EQ-5D-3L) representing (3^5) 243 possible health states as a generic health measure (2-4). In comparison, the newer EQ-5D five-level version (EQ-5D-5L) represents (5^5) 3125 possible health states resulting in increased sensitivity and reduced ceiling effects (16-22). Country-specific EQ-5D-5L preference-based value sets are available (https://euroqol.org/), with the value set for England (VSE) based on a composite Time Trade-Off (cTTO) and discrete choice experiment hybrid model (23-29). However, an independent quality assurance study led to NICE recommending the van Hout et al. cross-walk over the VSE (30-34). Therefore EQ-5D-5L preference-based values are cross-walked/mapped EQ-5D-3L United Kingdom (UK) value set scores based on the conventional TTO method (35). However, cross-walked scores have inherent concerns (e.g. predictive errors) and don't represent a direct value set for the EQ-5D-5L (36, 37). Analyses internationally comparing EQ-5D-5L and EQ-5D-3L value sets and alternative cross-walked scores suggest they estimate different preference-based values and subsequent QALYs (38-41).

Related to mental health, the EQ-5D measures' underlying health domains/items (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) have been argued to be more physical than mental health focussed, stimulating debate as to their appropriateness within mental health populations (10, 42-48). The 2010 Global Burden of Disease study estimated depression and anxiety disorders contribute a large portion of the total disability amongst all mental health and substance use disorders (49). Approximately 1/6 adults in England have a common mental health disorder (50). Mental health services and interventions have evolved to deal with care demand; for example, stepped-care within Improving Access to Psychological Therapies (IAPT) services in England and use of low-intensity interventions like Digital Mental Health Interventions (DMHIs) which require appropriate cost-effectiveness evidence (51-54). For reimbursement agencies like NICE, alternative preference-based measures can be rationalised based on aspects such as psychometric performance (4; p. 42), as suggested by Brazier and Deverill (55). EQ-5D measures' psychometric results offer better support in common (e.g. anxiety and depression) compared to severe (e.g. schizophrenia and bipolar disorder) mental health disorders (44-47, 56). The Recovering Quality-of-Life 20-item (ReQoL-20) and 10-item (ReQoL-10) versions are 'recovery-focussed quality-of-life' measures for mental health service users (57). A UK value set using the cTTO method has been developed to calculate QALYs from seven ReQoL-10 items: the ReQoL Utility Index (ReQoL-UI) representing (5^7) 78,125 possible health states (58). The ReQoL-UI's developers suggest it's arguably a more mental health focused generic measure relative to the more physical health focused EQ-5D measures (58). A psychometric analysis by Franklin, Enrique (59) in patients with anxiety and/or depression identified that compared to the EQ-5D-5L using the VSE or UK cross-walk, the ReQoL-UI had better construct validity with depression severity i.e. Patient-Health Questionnaire-9 (PHQ-9) score (60); whereby construct validity was assessed based on 'convergent' (e.g. correlation with the PHQ-9) and 'known-group' validity (e.g. assessing effect sizes between depression severity groupings; e.g. 'moderate' relative to 'mild' severity). However, the EQ-5D-5L preference-based score was more responsive (based on assessing standardised response means) and had better construct validity with anxiety severity i.e. Generalised Anxiety Disorder-7 (GAD-7) score (59, 61, 62). These results suggest that the two preference-based measures may systematically differ in how they measure anxiety and depression, with implications for the precision of QALY estimation (59).

We aim to explore the various QALY and subsequent cost-effectiveness estimates based on the EQ-5D-5L (VSE or cross-walk) or ReQoL-UI, when using alternative within-trial statistical methods based on a case study trial. Throughout we describe possible reasons for different QALY estimates based on the interaction between item scores, health state profiles, preference-based scores, mathematical and statistical methods chosen, with suggested implications for evaluating interventions within mental health services like IAPT and future research.

2. Methods

2.1. Data source

A parallel-groups, randomised waitlist-controlled trial examining the effectiveness and cost-effectiveness of internet-delivered Cognitive Behavioural Therapy (iCBT) for patients presenting with depression and anxiety was conducted at an established IAPT service (63, 64). Before 2:1 randomisation (intervention: 8-week waiting-list control), trial eligibility criteria was applied (Appendix S1). Trial inclusion criteria were people: (i) aged between 18-80 years; (ii) above clinical thresholds for depression (PHQ-9 ≥ 10) or anxiety (GAD-7 ≥ 8) (60-62), and (iii) suitable for iCBT (i.e. willing to use iCBT, internet access). The structured Mini International Neuropsychiatric Interview 7.0.2 (M.I.N.I.), administered by telephone by Psychological Wellbeing Practitioners (i.e. clinicians trained to deliver low-intensity support) established the presence or absence of a primary diagnosis of depression or anxiety disorder at baseline (65). NHS England Research Ethics Committee provided trial ethics approval (REC Reference: 17/NW/0311). The trial was prospectively registered: Current Controlled Trials ISRCTN91967124.

The trial is completed with the protocol and main results published showing that iCBT produced statistically significant improvements in depression (PHQ-9) and anxiety (GAD-7) severity compared to wait-list controls at 8-weeks, with further statistically significant intervention-group improvements from 8-weeks up to 12-months (63, 64). Over 8-weeks the probability of cost-effectiveness was 46.6% < £30,000 per EQ-5D-5L cross-walk-based QALY as the NICE reference-case (64). VSE and ReQoL-UI results were not published given NICE's VSE position and non-finalised ReQoL-UI at point of submission.

2.2. Preference-based measures

The EQ-5D-5L is a self-reported, generic health measure with five severity-levels over five dimensions/items (22). VSE and cross-walk score range: -0.285 or -0.594 to 1, respectively (25, 34).

The ReQoL-UI classification system is based on seven ReQoL-10 items: three positively (ReQoL-10 items: 5, 7, 10) and three negatively (ReQoL-10 items: 3, 6, 9) worded mental health items, and its one physical health item. These seven items cover seven themes of self-reported recovery-focused quality-of-life (58): autonomy; well-being; hope; activity; belonging and relationships; self-perception; physical health. The ReQoL-UI is described as having two overall dimensions: a mental health (six items) and a physical health (one item) dimension (58). ReQoL-UI score range: -0.195 to 1 (58).

2.3. Economic evaluation

This 8-week within-trial CUA focuses on the NICE reference case of cost-per-QALY from a health and social care perspective. As estimated QALYs are the main interest here, intervention (£94.63 per person) and other cost calculations are described elsewhere (64). We followed NICE guidelines, CHEERS checklist, and recommended methods for handling preference-based (utility), cost, and missing data using Stata version 15 and Microsoft Excel 2016 (4, 13-15, 30, 66-71).

2.3.1. *Calculating QALYs*

QALYs are calculated from preference-based scores using the total area-under-the-curve (AUC) method (15):

$$q_{jti} = \frac{(p_{j(t-1)i} + p_{jti})}{2} \delta_t \qquad \text{(Eq. 1)}$$

Whereby: p, preference-based score; i, an individual; t, time (i.e. baseline, t=0). For each group j (j=0, control; j=1, intervention), the consecutive time measures are added, averaged and then re-scaled (δ) for the percentage

of a year that *t* and *t-1* cover i.e. 0.15 for 8 weeks. From Eq. 1, total QALYs (Q) for each individual's trial duration are the summation of QALY calculations for each follow-up time-point starting at t=1:

$$Q_{ji} = \sum_{t=1}^{T} q_{jti} \qquad \text{(Eq. 2)}$$

Preference-based scores at baseline (t=0) and 8-week (t=1) are reported alongside subsequent QALY estimates for both trial-arms, and from 8-weeks to 12-months (t=5) for the intervention-arm only.

2.3.2. *Statistical analyses*

Analyses included complete-cases (CC) and with missing cases imputed based on multiple-imputation (MI) by chained equation using predictive mean matching, drawing inference from a pool of 10 donors (k-nearest neighbors = 10) thus avoiding predicting missing values outside the plausible/observed range (67, 72). The MI method was chosen post-hoc once the mechanism for missingness was deemed to be missing-at-random (MAR) based on logistic regression which identified baseline gender, GAD-7 caseness, work and social adjustment scale, and IAPT Phobia scale scores as predictors of missingness (13-15, 73). VSE, cross-walk, ReQoL-UI, and future cost missing cases at all follow-up time-points were imputed. Number of imputed datasets was based on percentage of missing CC data across all time-points in the intervention-arm (m = 43) (13, 74). Rubin's rule was applied when estimating MI analyses' means and standard errors of the mean (SEM) (75, 76).

Baseline-adjusted QALYs are estimated using baseline preference-based values and trial-arm as covariates within two independent regression models: ordinary least squares (OLS) and seemingly-unrelated regression (SUR), the latter accounting for the bivariate relationship between costs and QALYs (15, 77, 78). Incremental mean-point estimates of trial-arm differences (i.e. intervention minus control) related to mean costs over mean QALYs are used to estimate incremental cost-effectiveness ratios (ICERs).

Bootstrapping was used to calculate bootstrapped 95% confidence intervals (bCIs) and SEMs (bSEM) around costs and QALYs, and for plotting cost-effectiveness acceptability curves (CEACs). CC and MI analyses involved 5000 or 21500 (i.e. 500 nested within imputed datasets: m = 43) bootstrapped iterations, respectively (67). CEACs present the probability of intervention cost-effectiveness compared to control across a range of cost-effectiveness thresholds e.g. NICE's £20,000-£30,000 per QALY (4). CC analyses bCIs are bias corrected and accelerated (95% BCa CIs) which corrects for the bias and skewness in the distribution of bootstrap estimates, which is methodologically complicated for MI datasets when jackkniffing; therefore, Percentile method bCIs (95% bCIs) are used to reflect value coverage across bootstrapped MI datasets (13, 76).

Additional analyses exploring the interaction between estimated QALYs, preference-based scoring algorithms and item scores are described in the online Supplementary Appendices.

3. Results

3.1. Descriptive statistics

Overall, 361 people were randomised (241 intervention: 120 control): 71.5% were female, with a mean age of 33 years (range: 18-74). Baseline M.I.N.I diagnosis: 52%, major depressive disorder; 64%, anxiety disorder; 36%, both. The CC and MI analysis included 282 (194: 88) and 352 (236: 116) participants, respectively. Appendices S1-3 includes a Consort diagram, further demographic details, and measure completeness statistics.

3.2. Preference-based scores

Table 1 provides preference-based score descriptive statistics for observed cases at baseline across and within trial-arms. The cross-walk suggests this patient sample has the lowest, and the ReQoL-UI suggests the highest, mean preference-based health status at baseline. The EQ-5D-5L suggests this patient population is less heterogeneous than the ReQoL-UI, categorising 355 participants into 111 unique health state profiles (UHSPs); whereas the ReQoL-UI categorises 353 participants into 319 UHSPs. Relatedly, each ReQoL-UI UHSP is accompanied by its own unique preference-based score (UPBS). In comparison, 111 UHSPs are quantified by 100 VSE UPBSs and 105 cross-walk UPBSs, because some health states are represented by the same preference-based score (see Table 1).

Figure 1 shows kernel density estimates for the CC analyses preference-based scores at baseline and 8-weeks, as plotted on a graph within and across trial-arms; the change in score over this 8-week period is also presented. Figure 1 shows the VSE's distribution is 'smoother' than for the cross-walk, but not the ReQoL-UI, which is partly due to the number of UHSPs and UPBSs represented by each measure. Smoother in this context implies a broader distribution of scores across the score range resulting in less clustering and lower density around specific score ranges dependent on the pre-specified bandwidth (i.e. 0.02 for Figure 1). However and particularly at baseline, the ReQoL-UI presents higher density at the upper end of the scale (e.g. >0.7) compared to the cross-walk or VSE, which can relatively restrict ability for greater ReQoL-UI improvement post-baseline. Relatedly in the intervention-arm, the ReQoL-UI's high central density just above zero for 8-week score change is similar to the VSE and cross-walk, but the VSE and cross-walk have a broader distribution and additional peaks (e.g. >0.15) which contributes to a greater mean change.

Figure 2 presents MI mean and 95% CIs preference-based scores across all data collection time-points, and up to 8-weeks in Table 2. These results suggest crosswalk-based health is poorer than that estimated using the VSE or ReQoL-UI, which are more similar with each other than the cross-walk (Figure 2). The ReQoL-UI suggests that over 8-weeks the mean difference in preference-based health between trial-arms decreases, whereas the EQ-5D-5L suggested it increased with implications for estimating incremental QALYs. In the intervention-arm, a statistically significant difference with baseline preference-based scores is achieved by 8-weeks for the EQ-5D-5L but not until 3-months for the ReQoL-UI; this 3-month period represents the natural treatment timeframe in the intervention-arm not captured by the 8-week comparative trial-period nor the incremental QALY estimates.

3.3. Incremental results

Table 2 indicates the cross-walk produces the largest incremental QALY difference between trial-arms over 8-weeks; although, the ReQoL-UI produces more incremental QALYs than the VSE suggesting the opposite to the change in preference-based scores (Figure 2 and Table 2). This is because baseline imbalances are not accounted for across the individuals' total AUC calculations (Eq. 1), with regression-based adjustment recommended over individual-level adjustment as part of the AUC calculation (78, 79). Regression-based baseline-adjustment using the total AUC takes into account baseline imbalances in preference-based scores as well as the phenomenon that those individuals with preference-based scores that are lower or higher than the mean at baseline will usually experience a respectively higher or lower improvement at follow-up. Therefore because of the baseline imbalance and greater variation between the two-arms in the ReQoL-UI, when a baseline-adjustment is statistically applied the mean incremental difference in QALYs between the two arms is smaller than without the baseline-adjustment (78).

Table 3 and Figure 3 show that across both CC and MI unadjusted analyses, EQ-5D-5L and ReQoL-UI suggest iCBT is cost-effective < £30,000 per QALY (probability range: 54%-64%). Baseline-adjusted QALY results are contrary to the aforementioned, whereby for the same MI analyses the ReQoL-UI suggests the highest ICER (£1,252,542) relative to the cross-walk's lowest 'cost-effective' ICER (£27,684). When accounting for baseline-adjusted QALYs across CC and MI analyses, probability of cost-effectiveness < £30,000 per QALY ranged from 6% (CC ReQoL-UI baseline-adjusted QALY) up to 58.9% (CC cross-walk baseline-adjusted QALY). The largest change in probability of cost-effectiveness when moving from unadjusted to baseline-adjusted QALYs was for the ReQoL-UI in the MI analysis which dropped from 60.9% to 7.4% - an absolute decrease of 53.5%. Baseline-adjusted costs and SUR results are presented in Appendix S4.

The change in EQ-5D-5L and ReQoL-UI item-level scores are described in Appendix S5. To summarise, the EQ-5D-5L's cost-effectiveness results seems to be driven by the 'usual activities' and 'anxiety/depression' items, with the intervention-arm having better outcomes on average than the control-arm across all EQ-5D-5L domains. However, ReQoL-UI's item results were more varied, with the control-group having better outcomes on average than the intervention-arm across three (belonging and relationship; physical activity; self-perception) of its seven items, influencing the incremental ReQoL-UI results and subsequent QALY estimates.

4. Discussion

This study supports current empirical evidence that value sets like the VSE and cross-walked scores produce different QALYs even when from the same classification system (38, 39, 41, 80). We found that the VSE preference-based scores were more similar to those from the ReQoL-UI than the cross-walk. This meant the VSE and ReQoL-UI produced similar unadjusted QALYs. These similarities disappeared when statistically accounting for baseline preference-based scores, given that the control-group improved more on average than the intervention-group over 8-weeks on the ReQoL-UI – a result not mirrored on the EQ-5D-5L nor the trial's condition-specific (GAD-7 and PHQ-9) measures (64). This meant the ReQoL-UI had a lower probability of the intervention being cost-effective than the VSE or cross-walk: a decision-maker is unlikely to consider implementing iCBT based on these ReQoL-UI results, but might when using the cross-walk results. These differences stem from the analyses conducted (e.g. CC Vs. MI; unadjusted Vs. baseline-adjusted) and the measures themselves.

4.1. Exploring why the ReQoL-UI and EQ-5D-5L produce different QALYs

The different preference-based scores produced by the ReQoL-UI and EQ-5D-5L stem from aspects such as the content and size of their classification systems, the methods and populations used to value health states, and how their underlying preference-based scoring algorithms are constructed.

The ReQoL-UI can quantify a larger number of health states than the EQ-5D-5L (i.e. 78,125 Vs. 3,125), suggesting our study sample are more heterogeneous by categorising them into almost three times more health state profiles than the EQ-5D-5L. This categorisation stems from responses at the item-score level which indicated more response variability for the ReQoL-UI than the EQ-5D-5L (see Appendix S5). The ability to categorise population samples into more health states should permit the measure to be more sensitive to change in generic health status, as long as that change is represented by the measure's items and preference-based score.

As far as the current authors are aware, there is only one published psychometric assessment of the EQ-5D-5L and ReQoL-UI; a study conducted by the current authors using the same data source as this article specifically to inform the associated within-trial economic evaluation (59). This psychometric analysis suggests the ReQoL-UI has poorer responsiveness to change in GAD-7 anxiety or PHQ-9 depression severity than the EQ-5D-5L, which will have contributed to the smaller incremental QALY gains observed in this within-trial economic evaluation. Additionally, although the EQ-5D-5L was identified as having better construct validity with GAD-7 anxiety severity than the ReQoL-UI, the ReQoL-UI had better construct validity with PHQ-9 depression severity. The items that drove these construct validity results, particularly for the EQ-5D-5L, were the same items for which we identified a statistically significant difference between trial-arms over 8-weeks (e.g. 'anxiety/depression' and 'usual activities') as shown in Appendix S5 (59).

The ReQoL-UI has some perceived benefits over the EQ-5D-5L in mental health populations, including the ability to represent a larger number and variety of mental health states with better depression construct validity. However, in the MI analysis for example, the incremental ReQoL-UI baseline-adjusted QALYs were minimal (<0.0001) compared to those estimated from the VSE (0.0023) or cross-walk (0.0031); a result stemming in part from the ReQoL-UI's poorer responsiveness (particularly over 8-weeks). This is an unexpected result given that we would expect the ReQoL-UI to be more responsive given its mental health focus and classification system.

The psychometric analysis though only partly explains the different QALY estimations. Also influencing the result is that the ReQoL-UI's preference-based score is based on a "random effects model consisting of a quadratic specification of ɵ (newtheta) with interaction terms for ɵ and levels 3, 4, and 5 of physical health[sic]" (58). In other words, the physical health item/dimension has a direct interaction with the mental health dimension (ɵ) within the ReQoL-UI preference-based scoring algorithm. This is practically and conceptually different to how the EQ-5D value sets are scored with implications for the derived preference-based score. It is important that researchers currently/considering using the ReQoL-UI are aware of this interaction and associated rationale as described by Keetharuth A.D., Rowen D. (58). It is our hypothesis that the interaction with the physical health item contributed to the responsive statistics identified by the previous psychometric analysis and why the control group improved more on average than the intervention group in this IAPT-based within-trial analysis, as discussed further in section 4.2 and Appendix S6 (59).

4.2. Implications for mental health services, users, and research

The trial context is important for interpreting our results. IAPT Step 2 focusses on specific mental health populations and interventions; i.e. common mental health conditions that could benefit from low intensity therapies as brief psychological interventions (e.g. DMHI, Bibliotherapy) offered with support from clinicians (81). Furthermore, IAPT standards of patient recovery focus on symptom improvement, where 'recovery' is defined as moving from 'caseness' (PHQ-9≥10; GAD-7≥8) to 'no caseness' (54).

The ReQoL-UI psychometrics and within-trial results are potentially representative of its intended 'recovery-focussed' construct, which is different to 'recovery' as operationalised by IAPT. Such symptomatic changes seem to be captured in part by the EQ-5D-5L dimensions of 'usual activities' and 'anxiety/depression' which drive our within-trial results (Appendix S5). In comparison, the ReQoL-UI is developed from a conceptual framework of personal recovery in mental health, which is more focused on improving long-term wellbeing through self-management and having personally meaningful life goals, therefore expanding beyond the traditional symptom-based recovery paradigm (57, 82-85). As IAPT performance metrics are, in part, symptom-based recovery with a focus on mental health, previous psychometric results suggest the EQ-5D-5L captures these aspects better for anxiety severity and with greater responsiveness than the ReQoL-UI, and this is reflected in our IAPT-based within-trial economic evaluation results (59).

Additionally as mentioned in section 4.1, the ReQoL-UI's preference-based scoring algorithm includes a physical health interaction term with the mental health domain; this type of interaction term is not used in the EQ-5D measures' preference-based value set scoring algorithms. Step 2 IAPT patients are referred on the basis of suffering from acute depression and/or anxiety symptomology, with improvements in physical health not being a key purpose of the service. In this trial's IAPT-based population the majority of participants reported baseline physical health as 'no problem' or 'slight problem', with the majority not moving from this baseline state (Appendix S5). The interaction term in the ReQoL-UI's preference-based scoring algorithm means that because the majority of the study sample have no/slight problems with baseline physical health from which there is no change over the trial period, there is subsequently restricted ability for the ReQoL-UI's preference-based score to change, even if there are changes across the mental health domain. This will have influenced the ReQoL-UI's responsiveness, but also incremental QALY estimates; particularly given the control-arm randomly had more people who reported worse physical health at baseline and had a higher mean improvement in physical health over 8-weeks than the intervention-arm. Compared to the ReQoL-UI, for the EQ-5D values sets an interaction term is not imposed between the physical and mental health items allowing more independence between items in the preference-based scoring algorithm – this aspect is explored further in Appendix S6.

In different mental health settings (e.g. hospital outpatients), patient populations (severe mental health disorders), with different intervention types (high-intensity interventions), these psychometric and within-trial results could be different. Further research is warranted including to what extent various mental health interventions, from medication to DMHIs, are intended to promote symptomatic or personal recovery, as well as physical health, which itself could dictate if the EQ-5D-5L or ReQoL-UI may be the more appropriate preference-based measure to estimate cost-effectiveness. Further exploratory analysis of the ReQoL-UI is warranted before it is used to guide resource-allocation decision-making, particularly as a complement or substitute to the EQ-5D-5L. Additionally, EuroQol's Blog provides updates for its new health and wellbeing instrument (EQ-HWB) which should be considered for future research (86).

4.3. Limitations

The 8-week between trial-arm analyses limited the ability to assess incremental QALYs over a longer time-horizon. Common mental health disorder trials rarely exceed 12-month follow-up, with most follow-up periods aligning with when clinical change is most likely to be observed following treatment: between 8-12 weeks (87-89). The lack of longer-term data also limits the ability and/or reliability of conducting extrapolated or modelling-based analyses over an even longer time-horizon. A systematic review of DMHI economic evaluations stated that 54 of 66 included papers did not explore the results beyond trial end-points: "lack of longer-term modelling is likely to be due to, in part, the lack of reliable data about the long term performance of

DMHIs" (51). These data-driven limitations suggest longer-term comparative trial follow-ups are needed whenever possible with statistical methods as secondary options (14, 90, 91).

The VSE has suggested complications beyond what our analysis explores, with a new UK valuation study underway (31-33, 59, 92). However, as an imperfect direct value set for the EQ-5D-5L relative to the cross-walk which represents the EQ-5D-3L UK value set, it is still useful and informative for this exploratory analysis.

5. Conclusion

These results indicate the importance of reflecting on a preference-based measure's whole design before using it for economic evaluation, aspects of which can be revealed by conducting psychometric analyses, as on QALY face value it is difficult to wholly understand why different preference-based measures produce different QALYs. These differences stem from mathematical and statistical methods used and the preference-based measure itself, which need to be considered holistically to understand any subsequent QALY and cost-effectiveness estimates before suggesting to decision-makers if an intervention is 'cost-effective' or not based on such evidence.

References

1. Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the economic evaluation of health care programmes. Oxford University Press, 2015.
2. 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.
3. Kennedy-Martin M, Slaap B, Herdman M, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. Eur J Health Econ. 2020.
4. NICE. Guide to the methods of technology appraisal. In: National Institute for Health and Care Excellence (NICE), ed. London, 2013.
5. Brazier J, Ratcliffe J, Saloman J, et al. Measuring and valuing health benefits for economic evaluation. Oxford University Press, 2016.
6. Brazier J, Ara R, Rowen D, et al. A review of generic preference-based measures for use in cost-effectiveness models. Pharmacoeconomics. 2017; 35: 21-31.
7. Brazier J, Rowen D. NICE DSU Technical Support Document 11: Alternatives to EQ-5D for Generating Health State Utility Values In: NICE DSU, ed., NICE DSU Technical Support Document (TSD). Sheffield, 2011.
8. Rowen D, Brazier J, Ara R, et al. The role of condition-specific preference-based measures in health technology assessment. Pharmacoeconomics. 2017; 35: 33-41.
9. Versteegh MM, Leunis A, Uyl-de Groot CA, et al. Condition-specific preference-based measures: benefit or burden? Value Health. 2012; 15: 504-13.
10. Brazier J. Measuring and valuing mental health for use in economic evaluation. J Health Serv Res Policy. 2008; 13: 70-75.
11. Lancsar E, Gu Y, Gyrd-Hansen D, et al. The relative value of different QALY types. J Health Econ. 2020: 102303.
12. Weinstein MC. A QALY is a QALY is a QALY—or is it? J Health Econ. 1988; 7: 289-90.
13. Faria R, Gomes M, Epstein D, et al. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. Pharmacoeconomics. 2014; 32: 1157-70.
14. Franklin M, Lomas J, Walker S, et al. An educational review about using cost data for the purpose of cost-effectiveness analysis. Pharmacoeconomics. 2019: 1-13.
15. Hunter RM, Baio G, Butt T, et al. An educational review of the statistical issues in analysing utility data for cost-utility analysis. Pharmacoeconomics. 2015; 33: 355-66.
16. Golicki D, Niewada M, Karlińska A, et al. Comparing responsiveness of the EQ-5D-5L, EQ-5D-3L and EQ VAS in stroke patients. Qual Life Res. 2015; 24: 1555-63.
17. Buchholz I, Thielker K, Feng Y-S, et al. Measuring changes in health over time using the EQ-5D 3L and 5L: a head-to-head comparison of measurement properties and sensitivity to change in a German inpatient rehabilitation sample. Qual Life Res. 2015; 24: 829-35.
18. Janssen MF, Birnie E, Haagsma JA, et al. Comparing the standard EQ-5D three-level system with a five-level version. Value Health. 2008; 11: 275-84.
19. Pickard AS, De Leon MC, Kohlmann T, et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007; 45: 259-63.
20. Scalone L, Ciampichini R, Fagiuoli S, et al. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. Qual Life Res. 2013; 22: 1707-16.
21. Golicki D, Zawodnik S, Janssen MF, et al. Psychometric comparison of EQ-5D and EQ-5D-5L in student population. Value Health. 2010; 13: A240-A40.
22. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011; 20: 1727-36.
23. 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: 445-53.
24. Stolk E, Ludwig K, Rand K, et al. Overview, update, and lessons learned from the International EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. Value Health. 2019; 22: 23-30.
25. Devlin NJ, Shah KK, Feng Y, et al. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Econ. 2018; 27: 7-22.
26. 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: 348-61.
27. Janssen BM, Oppe M, Versteegh MM, et al. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013; 14: 5-13.
28. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, et al. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care. 2017; 55: e51.

29. Rowen D, Brazier J, Van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. Med Decis Making. 2015; 35: 328-40.
30. NICE. Position statement on use of the EQ-5D-5L valuation set for England (updated November 2018). London: National Institute for Health and Care Excellence (NICE), 2018.
31. Hernández-Alava M, Pudney S, Wailoo A. Quality review of a proposed EQ-5D-5L value set for England. EEPRU report [online]. 2018.
32. Norman R, Olsen JA. Competing Views on the English EQ-5D-5L Valuation Set. Value Health. 2020; 23: 574-75.
33. van Hout B, Mulhern B, Feng Y, et al. The EQ-5D-5L Value Set for England: Response to the "Quality Assurance". Value Health. 2020.
34. van Hout B, Janssen M, Feng Y-S, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012; 15: 708-15.
35. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997: 1095-108.
36. Mukuria C, Rowen D, Harnan S, et al. An Updated Systematic Review of Studies Mapping (or Cross-Walking) Measures of Health-Related Quality of Life to Generic Preference-Based Measures to Generate Utility Values. Appl Health Econ Health Policy. 2019: 1-19.
37. Longworth L, Rowen D. NICE DSU technical support document 10: the use of mapping methods to estimate health state utility values. Sheffield: Decision Support Unit, ScHARR, University of Sheffield. 2011: b4.
38. Gerlinger C, Bamber L, Leverkus F, et al. Comparing the EQ-5D-5L utility index based on value sets of different countries: impact on the interpretation of clinical study results. BMC Res Notes. 2019; 12: 1-6.
39. Mulhern B, Feng Y, Shah K, et al. Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets. Pharmacoeconomics. 2018; 36: 699-713.
40. Hernández Alava M, Wailoo A, Grimm S, et al. EQ-5D-5L versus EQ-5D-3L: the impact on cost effectiveness in the United Kingdom. Value Health. 2018; 21: 49-56.
41. Wailoo A, Alava MH, Pudney S, et al. An International Comparison of EQ-5D-5L and EQ-5D-3L for Use in Cost-Effectiveness Analysis. Value Health. 2021.
42. Franklin M. Cost Utility Analysis. In: Razzouk D, ed., Mental Health Economics: Springer, 2017.
43. Razzouk D. Mental Health Economics: The Costs and Benefits of Psychiatric Care. Springer, 2017.
44. 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.
45. 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: 236-43.
46. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics. 2015; 33: 1137-54.
47. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. Eur J Health Econ. 2018; 19: 557-70.
48. Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. Health Technol Assess. 2014.
49. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet. 2013; 382: 1575-86.
50. McManus S, Bebbington P, Jenkins R, et al. Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey 2014: a Survey Carried Out for NHS Digital by NatCen Social Research and the Department of Health Sciences, University of Leicester. NHS Digital, 2016.
51. Jankovic D, Bojke L, Marshall D, et al. Systematic Review and Critique of Methods for Economic Evaluation of Digital Mental Health Interventions. Appl Health Econ Health Policy. 2020: 1-11.
52. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. Int Rev Psychiatry. 2011; 23: 318-27.
53. Gyani A, Shafran R, Layard R, et al. Enhancing recovery rates: lessons from year one of IAPT. Behav Res Ther. 2013; 51: 597-606.
54. NHS Digital. A guide to IAPT data and publications. NHS Digital, 2021.
55. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ. 1999; 8: 41-51.
56. Brazier J, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004; 13: 873-84.
57. 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.
58. Keetharuth A.D., Rowen D., Bjorner J., et al. Estimating a Preference-Based Index for mental health from the Recovering Quality of Life (ReQoL) measure: Valuation of ReQoL-UI. Value Health. 2020.

59. Franklin M, Enrique A, Palacios J, et al. Psychometric assessment of EQ-5D-5L and ReQoL measures in patients with anxiety and depression: construct validity and responsiveness. Qual Life Res. 2021.
60. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16: 606-13.
61. Kroenke K, Spitzer RL, Williams JB, et al. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med. 2007; 146: 317-25.
62. Spitzer RL, Kroenke K, Williams JB, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006; 166: 1092-97.
63. Richards D, Duffy D, Blackburn B, et al. Digital IAPT: the effectiveness & cost-effectiveness of internet-delivered interventions for depression and anxiety disorders in the Improving Access to Psychological Therapies programme: study protocol for a randomised control trial. BMC Psychiatry. 2018; 18: 59.
64. Richards D, Enrique A, Eilert N, et al. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. NPJ Digit Med. 2020; 3: 1-10.
65. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry. 1998.
66. Microsoft Corporation. Microsoft Excel 2016. 2016.
67. Leurent B, Gomes M, Faria R, et al. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. Pharmacoeconomics. 2018; 36: 889-901.
68. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC,, 2017.
69. Ramsey S, Willke R, Briggs A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. Value Health. 2005; 8: 521-33.
70. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials II—an ISPOR Good Research Practices Task Force report. Value Health. 2015; 18: 161-72.
71. Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS)—explanation and elaboration: a report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. Value Health. 2013; 16: 231-50.
72. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. 2014; 14: 1-13.
73. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, 2002.
74. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011; 30: 377-99.
75. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons, 2019.
76. Burton A, Billingham LJ, Bryan S. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. Clin Trials. 2007; 4: 154-61.
77. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Econ. 2004; 13: 461-75.
78. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ. 2005; 14: 487-96.
79. Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. Health Econ. 2004; 13: 1203-10.
80. Hernández Alava M, Pudney S, Wailoo A. Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from an English Population Study., EEPRU Report: University of Sheffield & University of York, 2020.
81. Bennett-Levy J, Farrand P, Christensen H, et al. Oxford guide to low intensity CBT interventions. Oxford University Press, 2010.
82. Leamy M, Bird V, Le Boutillier C, et al. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. Br J Psychiatry. 2011; 199: 445-52.
83. Shepherd G, Boardman J, Rinaldi M, et al. Supporting recovery in mental health services: Quality and outcomes. Implementing Recovery Through Organisational Change (ImROC), 2014.
84. Slade M, Longden E. Empirical evidence about recovery and mental health. BMC Psychiatry. 2015; 15: 1-14.
85. Onken SJ, Craig CM, Ridgway P, et al. An analysis of the definitions and elements of recovery: a review of the literature. Psychiatr Rehabil J. 2007; 31: 9.
86. EuroQol. EuroQol is developing a new instrument – The EQ-HWB (Blog post). EuroQol, 2020.
87. Ramsberg J, Asseburg C, Henriksson M. Effectiveness and cost-effectiveness of antidepressants in primary care: a multiple treatment comparison meta-analysis and cost-effectiveness model. PLoS One. 2012; 7: e42003.
88. Annemans L, Brignone M, Druais S, et al. Cost-effectiveness analysis of pharmaceutical treatment options in the first-line management of major depressive disorder in Belgium. Pharmacoeconomics. 2014; 32: 479-93.
89. NICE. Depression in adults: recognition and management. In: National Institute for Health and Care Excellence (NICE), ed., NICE clinical guideline. London, UK, 2009.

90. Richardson J, Khan MA, Iezzi A, et al. Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. Med Decis Making. 2015; 35: 276-91.
91. Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. Oxford University Press, 2006.
92. EuroQol. New UK EQ-5D-5L Valuation Study (Blog post). EuroQol, 2020.