Towards a patient-reported summary score for EQ-5D (EQ-PRSM)

Janssen MF[1,2], Finch AP[2], Bonsel GJ[2,3]

1 – Section Medical Psychology and Psychotherapy, Department of Psychiatry, Erasmus MC, Rotterdam, The Netherlands
2 – EuroQol Research Foundation, Rotterdam, the Netherlands
3 – Department of Public health, Erasmus MC, Rotterdam, The Netherlands

Abstract

Objectives
EQ-5D is increasingly being used outside the context of health technology assessment, e.g. as patient-reported outcome measure in patient registries, in population health studies and in personalized medicine. For purposes other than economic evaluation, there is no clear rationale for using value sets to summarize EQ-5D profile data. There use of values for descriptive purposes illustrates the need for a global, single summary measure based on self-reported data, be it from patients or members of the general public. In this study we explored two theory-driven approaches to obtain summary scores for EQ-5D for non-economic applications.

Methods
Three large multi-country datasets were used: the "crosswalk" dataset, conducted in seven patient groups (cardiovascular disease, respiratory disease, depression, diabetes, personality disorders, arthritis and stroke) across five countries (N=2,707), and two multi-country general population datasets: QOLIBRI (N=10,172), and MIC (N=8,022). Two methodological approaches were explored. The first derives patient-reported summary scores from dimension-specific rating scales (RS) for each EQ-5D-5L dimension. This RS approach is based on existing psychometric methodology, allowing for a more refined assessment of the underlying position of the level responses for each dimension. Mean and median RS scores were used to calculate additive and (EQ VAS) weighted RS summary scores. The second approach fits Item Response Theory (IRT) models to EQ-5D dimension responses. Models tested were Rasch, a unidimensional 2 parameter graded response model (GRM) and a multidimensional 2 parameter GRM model (requiring items beyond EQ-5D). Location parameters and theta values were used to generate the summary scores.

Results
For the RS approach the EQ-5D-5L dimension – RS response pairs occasionally showed inconsistent responses (scale reversal). Mild and strong exclusion criteria were subsequently adopted, excluding 10.6% and 16.5% respondents, respectively. Mild versus strong criteria, or means versus medians, produced similar summary score results. We selected the mild exclusion criteria, mean-based models as preferred option. The EQ VAS weighted algorithm resulted in a smoother modelled summary score distribution, compared to a simple additive approach. For the IRT approach: the unidimensional IRT GRM model resulted in a low discrimination parameter for one item i.e. anxiety/depression. The 2 parameter GRM multidimensional model reported good fit statistics, well ordered categories and large and significant discrimination parameters, producing an alternative to the RS approach to develop EQ-5D summary scores.

Conclusion
This study presents two alternative methods to arrive at patient-reported summary scores for EQ-5D-5L, for use in population health and health systems applications. The methodological approaches can be applied to any scalable multidimensional health instrument. In a follow up study we will empirically test the different summary scores for measurement properties.

Introduction

One of the greatest strengths of the EQ-5D family of instruments is the available of country-specific value sets, facilitating the calculation of quality-adjusted life years (QALYs) that are used to inform economic evaluations of health care interventions or policies on health [1]. EQ-5D value sets are specifically designed for that purpose, which is reflected both in the underlying methodology (based on stated preferences) and whose preferences are sought (by convention the general public). For purposes other than the estimation of QALYs, there is no clear rationale for using any value set to summarize profile data [2]. However, it seems that for these ('non-QALY') purposes values still are often used as a means of summarizing scores for the five dimensions of EQ-5D [4-6].

EQ-5D is increasingly being used in applications outside of health technology assessment (HTA), e.g. as patient-reported outcome measure (PROM) in population health and health system applications. In population health assessment EQ-5D has been used to assess determinants of health and health inequalities [7-9]. EQ-5D has been used on different levels of health systems in order to monitor, evaluate and improve quality of care, but also to enhance patient-centered care [10, 11]. For these purposes, there is no need to use preference-based (index) values. The use of index values for descriptive purposes demonstrates a clear need, however, for a single summary score based on the responses of the respondents themselves, be it patients or members of the general public [5, 12, 13].

We conducted a proof of principle study on the development of a patient-reported summary score (PRSM) for EQ-5D for non-economic purposes. We selected two different theory-driven approaches, both of which seem suitable for the purpose. The first derives summary scores from dimension-specific rating scale responses for each of the five EQ-5D-5L dimensions, to which an empirical (non-preference) weighting is added. The second approach fits Item Response Theory (IRT) models to the EQ-5D dimension responses, assuming one (or more) latent trait(s), which can combined into a single summary score.

Methods

Data sources

Three available large international datasets were used for the current study. Two datasets by design included the EQ-5D-5L and a set of five dimension-specific rating scales (on a 0-100 scale), one for each of the five EQ-5D dimensions, following an earlier study design [14]. The first dataset was resulting from the EQ-5D-5L study for interim value sets (the "crosswalk" study), conducted in eight patient groups (cardiovascular disease, respiratory disease, depression, diabetes, liver disease, personality disorders, arthritis and stroke) and a student cohort in Denmark, Italy, the Netherlands, Poland and the United Kingdom [15, 16]. Data were collected in face-to-face paper and pencil interviews, resulting in a total sample size of 3,919. All sub-samples included the dimension-specific rating scales expect for the student cohort and the liver disease sample (Italy). The second, so-called QOLIBRI dataset was collected in 2017 as a web-based survey to members of the general public aged 18 to 70 years from three European countries (the United Kingdom, Italy and the Netherlands) [9, 17]. The respondents were selected in such a way that they were representative of the population aged 18 to 70 in the countries with respect to age, sex, and educational level. A total number of 11,759 respondents filled in the questionnaire (Italy: 3,549 respondents; Netherlands 3,564

respondents; UK 4,646 respondents). The third dataset was the multi-instrument comparison (MIC) dataset: an online population health survey conducted in six countries (Australia, Canada, Germany, Norway, the United Kingdom, and the United States) where quota sampling ensured similar sociodemographic characteristics between countries aimed at population representativeness [18]. The final sample included 8,022 individuals, who completed the EQ-5D-5L and other health and wellbeing measures, the dimension-specific rating scales were not included.

Only respondents with no missing responses on the instruments included in the analysis were used. An overview of the sample characteristics of the datasets is shown in Table 1 and Appendix 1. Across population subgroups, % female respondents varied from 36% (liver disease) to 79% (students), the mean age varied from 22 (students) to 68 years (stroke), and mean EQ VAS (SD) ranged from 53 (26) for the stroke sample to 79 (16) for the student sample. Crosswalk and MIC datasets had an approximately equal distribution of gender (52% females), and age (mean 52 years). However, responders in the MIC dataset appeared generally healthier than responders in the crosswalk dataset, based on the self-reported health using the EQ-5D (Appendix 1).

Methodological approach

Two methodological approaches were applied. The first approach is based on existing psychometric methodology and creates patient-reported summary scores from dimension-specific rating scales (RS) for each of the five EQ-5D-5L dimensions. The RS approach allows for a more refined assessment of the underlying position of the level responses for each dimension. RS data were available from 8 crosswalk population groups and the QOLIBRI general population sample (Table 1). The second approach is based on Item Response Theory (IRT), applying unidimensional IRT (UIRT) models and multidimensional IRT (MIRT) to self-reported EQ-5D-5L data. For this proof-of-principle study, we used the optimum data(sets) available for each approach, and each method within each approach (Table 1). The resulting summary score models will therefore be of limited comparability.

RS approach: Theoretical background

Rating scales and visual analogue scales have their theoretical foundation in psychological theories in response to sensory stimuli, and have a long history in psychometric research. The method has been used extensively in health and health-related quality of life (HRQL) to assess a variety of constructs like pain, mood and functional capacity, among others [19-21].

From earlier research in comparing descriptive systems of varying granularity, we learned that the more crude the classification is in terms of number of levels, the higher the probability that biases occur [14, 22]. This is related to the 'true' location of the respondent or patient on the scale underlying the response options. The RS approach is a way to attempt to determine a more precise position on the scale, tailor made to each of the five dimensions. This approach, or similar approaches, have been used before [23, 24] and have proven to be useful in arriving at PRSM scores.

Our approach has a close relation to rather recent developments where the focus of outcomes, values and preferences is on the patient, through the development of patient values, patient-experience based values and patient-centered measures [25-27]. An example of this is the so-called patient-experience based value set approach [28, 29]. Our approach goes beyond currently applied methods of experience-based value sets, by measuring a more refined way of assessing patients'

health for each dimension, and exploring more advanced methods of establishing the final summary scores. Furthermore, we think experience-based values would not need to be limited to measurement in a general population sample, but could much better be focused on where the scores will actually be used: the patient. That is why our modeling dataset for our approach for the most part consists of a large international dataset for different patient groups with conditions of varying severity.

RS based approach: Empirical strategy and analysis

Only respondents with no missing responses on either EQ-5D-5L or RS were included, resulting in sample sizes of 3,187 (crosswalk dataset) and 10,172 (general population). RS scales consisted of horizontal hashmarked lines from 0 to 100 with corresponding numbers (0, 10, 20, …, 100). The descriptive anchors at each end of the scales were the same anchors as used in EQ-5D-5L: "no problems" (100) and "unable to/extreme problems" (0).

RS analyses were performed using Stata 16.1. We calculated mean and median RS scores for each of the levels for all five EQ-5D dimensions separately, for each population subgroup. Only average level scores were used with at least 10 observations. As we expected there to be differences in average RS scores between populations (e.g. caused by response heterogeneity or response shift) we calculated overall averages for the mean and medians by using equal weights for all 9 population groups. Mean and median scores for levels 1 and 5 were also used (as they can deviate from scores 100 and 0 respectively). Next, mean and median RS scores were transformed to a 0-20 range for each dimension, so that the final total PRSM score will result in a 0-100 range.

We adopted two weighting approaches to aggregate the five rating scale (RS) scores into a single summary score. The first approach applies equal weights, while the second approach applies weights derived from a regression method using the EQ VAS as dependent variable. To take the relative contribution of the different dimensions on the overall PRSM score into account, we performed ordinary least square regressions of the full EQ-5D-5L dimensions onto the EQ VAS for all 9 population groups separately. Overall average regression weights were calculated by using equal weights for all 9 population groups. Unstandardized b coefficients for the each dimension were used as weights to adjust for the importance of each of the dimensions, by multiplying each dimension level with the corresponding weight to arrive at the PRSM model. A linear transformation was performed to arrive at a 0-100 PRSM score.

Figure 1 provides a schematic overview of the RS approach for arriving at a PRSM score for EQ-5D-5L.

Item response theory approach: Theoretical background and analysis

IRT analyses were performed using M-Plus version 7©. For the main analysis, the pooled crosswalk dataset was used, to ensure a good spread of observed response across all levels and all dimensions. We fitted a 2 parameter (2PL) unidimensional Graded Response Model (GRM) [30] and a 1 parameter (1PL) Rasch model [31] via marginal maximum likelihood (MML) estimator.

As the sample size did not allow for assessing absolute model fit with this estimator, we assessed relative fit instead, using a maximum likelihood test. We also fitted the same 2PL and 1PL models using the limited information diagonally Weighted Least Square Estimator Mean and Variance corrected (WLSMV), which summarizes item responses into matrices of polychoric correlations prior

to fitting the 2PL or 1PL model. The WLSMV estimator provides two practical fit indexes for the 2PL and 1PL models: the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). Model fit was used to compare between models and to assess the general fit of a model. For the latter purpose, RMSEA was considered acceptable when 0.08 or less and good when 0.05 or less, and the CFI acceptable when 0.90 or more and good when 0.95 or more, as suggested in international guidelines [32-34]. Beyond this absolute fit assessment, the 2PL and 1PL models could be compared. Additional models were also estimated, including generalization of the 1PL model with freely estimated equal discriminations across items. Based on theoretical considerations and model fit results, the 2PL GRM is the preferred model, as further discussed below.

The 2PL GRM partitions the five EQ-5D-5L items Likert scales into a series of binary options (in other research contexts also referred to as cut-points), for which there are k-1 sub models per item, where k is the number of response options, here 5-1=4. Sub-models are then estimated on the cumulative data, producing a set of discrimination parameters (also commonly referred as alpha (a)) and a set of category specific thresholds (also referred as difficulty parameters, or b). Discrimination parameters are a measure of differential capability of the item i.e., how much the item is related to the latent trait and can therefore distinguish between individuals. Difficulty parameters represent the quantity of the latent trait required to have/associated with a probability of at least 50% to answer positively to a given response level. In that, they can be conceptualized as the level of latent trait from which each given EQ-5D dimension level is most likely to be chosen.

We used a standardized transformation of latent theta scores with mean 0 and variance 1, meaning all parameters for all response levels are on the same scale and expressed in terms of standard deviation from the mean.

The test the 2PL GRM model for the three assumptions the model relies on, we used Mokken scale analysis for polytomous items for monotonicity, fitted confirmatory factor analysis models for testing unidimensionality and investigated residual correlation matrices for local independence [31]. Recent research into the EQ-5D-5L using the MIC data showed it forms a moderate to strong Mokken scale and therefore is monotonic [35], and that is moderately unidimensional [36] albeit a multidimensional structure is preferable [37].

Based on these results and data constraints (absence of other measures in the crosswalk data), we first fitted a unidimensional 2PL GRM model using the crosswalk and MIC datasets. A multidimensional 2PL GRM model was then fitted (MIC dataset only) [38], “borrowing” items from other instruments. For the sake of parsimony and to maximize illustration of the principle, we specified a 3 factor IRT model with 4 items per factor (domain) only. The data offered to the multidimensional 2PL GRM model consisted of a selection of physical functioning (mobility, selfcare, activities and HUI ambulation), mental health (anxiety and depression, 15D distress, SF mental, AQoL frequency sadness) and pain (Pain/ discomfort, AQoL frequency of pain, HUI pain, AQol severity of pain) items. The EQ-5D was assumed to be related to 3 latent traits, physical functioning, pain and mental health, with variable number of items per latent trait. The items which complemented the EQ-5D were selected from the list of items loading on the same pain and mental health factors in previous work of Finch et al [37].

Different IRT model outputs were created. For the GRM models, we first report discrimination and difficulty parameters. We then use two post estimation methods to score the EQ-5D. The first

method uses the level of theta that best represented the individual who is most likely to choose each of the item response levels. These levels of theta were extrapolated post estimation using the item characteristics curve data of each individual item, i.e. to estimate the top of the curves. As it is not possible to identify the specific level of theta where the most representative individual for level 5 lies (as the probabilities are asymptotic towards the left end of the scale), we arbitrarily choose the theta level associated to the mid-point between the probability of 1 and the probability associated to the point in which level 5 becomes more likely than level 4. This corresponds to the expected theta value of the most representative individual for perfectly behaved items. For practicality (as theta values can be both positive or negative), difficulty coefficients can be rescaled to an arbitrary target range that ensures conformity with the relative differences in the logit values generated using the GRM model. The second method uses the IRT scoring mechanism called expected a posteriori [39-41]. This scoring generates a posterior probability curve based on Bayesian priors, by reshaping the probability curves as a function of both the prior distribution and the probability curve [42]. It allows in this way to generate theta scores, which can be conceptualized as the numerical value representing the placement of individuals on the latent trait. We regressed the latent theta values over EQ-5D dummies and reported beta coefficients, which represent the amount of decrease in latent trait associated with the level of the dummy variable compared with the reference case (best possible health). The regressions assume additivity of predictors and normal distribution of the theta trait. For UIRT models, there is a single latent trait, while for the MIRT model, three latent traits are regressed over the EQ-5D-5L dummies. We present the results of the MIRT regressions for completeness and to inform future research, albeit the issue of how to combine them still can be improved upon. For both approaches, theta values and b coefficients were rescaled so resulting PRSM scores were on a 0 – 100 scale.

Descriptive comparison of the scoring systems

The resulting EQ-PRSM models will be compared descriptively by dimension impact order, and visually using histograms for all 3,125 possible modelled scores.

Results

RS approach

After excluding respondents with missing responses on either EQ-5D-5L or RS, sample sizes were 2,707 (crosswalk dataset) and 10,172 (QOLIBRI general population). There were several level 5 categories with less than 10 observations for the various population groups (Appendix 1), ranging from 1 (anxiety/depression) to 7 (self-care). There were also a few level 4 and level 3 categories with <10 observations. When applying the RS approach we observed some inconsistent EQ-5D-5L dimension – RS response pairs. Apparently RS scales were reversed in some respondents. Mild and strong exclusion criteria were subsequently adopted. Using mild criteria, respondents were excluded when they scored level 1 or 2 on a certain dimension paired with an RS score of ≤30, or ≤20, respectively, or when they scored level 4 or 5 on a certain dimensions paired with an RS score ≥80 or ≥70, respectively. Strong criteria were applied by excluding respondent when they scored level 1 or 2 on a certain dimension paired with an RS score of ≤50, or ≤30, respectively, or when they scored

level 4 or 5 on a certain dimensions paired with an RS score ≥70 or ≥50, respectively. In total, 10.6% of respondents were excluded adopting mild criteria, and 16.5% applying strong criteria.

Level distributions by dimension for all datasets and population groups are shown in Appendix 1. There was considerable variation in mean RS scores, with an average difference of 13 RS points (not including the categories with less than 10 observations) (Appendix 1).

The effect of mild or strong exclusion criteria, and of mean-or median-based models was relatively small on the resulting PRSM models. Therefore, to take a conservative approach, we chose the mild criteria models as preferred option, and also opted for the mean-based RS score model. EQ VAS regression weights between dimensions varied considerably across population groups (Appendix 2). Overall, anxiety/depression and pain/discomfort showed the most severe weights, whereas self-care and mobility were the mildest. Usual activities varied the most between populations groups, sometimes showing the most severe weight (Asthma/COPD, personality disorder, stroke and other condition groups) and in one instance the mildest (cardiovascular disease). The average EQ VAS weights were 3.37 (mobility), 1.64 (self-care), 4.45 (usual activities), 4,51 (pain/discomfort), and 4.96 (anxiety/depression).

Table 2 shows the equally and EQ VAS weighted EQ-PRSM models based on the RS approach.

IRT approach

Using the crosswalk data for the main IRT analyses, the 1PL Rasch model (crosswalk dataset) reported a log likelihood of -19183.162, the 2PL model a log likelihood of -18475.634, and the two models differed in terms of 4 degrees of freedom i.e., parameters for the estimation of 4 additional discriminations in the 2PL model, which are constrained as equal in the Rasch 1PL model. The test was statistically significant, showing that the 2PL model should be preferred over the 1PL model. Comparison of the two models in terms of RMSEA and CFI, when using the WLSMV estimator, showed the 2PL model had a substantially better fit than the Rasch 1PL model, with a CFI of 0.988 versus 0.921 and a RMSEA of 0.149 versus 0.300. Of note, also for the 2PL model the model fit was not optimal.

Table 3 shows discrimination and difficulty parameters for the UIRT GRM 2PL model, using the crosswalk and the MIC datasets. The estimated discrimination parameter in the crosswalk dataset ranged between 4.873 for Selfcare to 0.885 for anxiety/depression, and from 4.506 for usual activities to 0.942 for anxiety/depression in the MIC data. This suggests that in both datasets, anxiety/depression is the least related to the latent trait measured, followed by pain/discomfort. Larger discriminations for usual activities in the MIC dataset compared to the crosswalk indicate that this concept is more closely related to health in this sample compared to the crosswalk sample. The difficulty parameter estimates were always lower in the MIC dataset compared to the crosswalk dataset, signaling that responders in this dataset required less health to respond to a higher response category. Table 3 also reports the discrimination and difficulty coefficients for the MIRT 2PL GRM model (MIC dataset). As it can be seen, the discrimination parameters of the MIRT model are substantially larger for pain/discomfort and anxiety/depression compared to the ones reported in the UIRT model, showing that the items are well represented by the new latent trait measured. Discrimination coefficients remain high for the other EQ-5D items.

Table 4 reports the most likely level of the latent trait using the first estimation method. As it can be seen, at increasing severity, problems were associated lower levels of theta, for all EQ-5D dimensions. Anxiety/depression level 5 registered a substantially lower level of theta compared to the other EQ-5D dimensions. This was not the case when using MIRT, where all EQ-5D dimensions, including anxiety/depression, covered a similar range of the latent traits measured. For the second method, table 4 reports the b coefficients of the regression of expected a posteriori theta over the EQ-5D dummy variables. For both datasets, using both models, coefficients were statistically significant and monotonically decreasing. When using UIRT coefficients for the anxiety/depression covered a substantially smaller range compared to the other EQ-5D dimensions. This was not the case for the MIRT.

Figures 2A-D reports the item characteristics curves for two selected EQ-5D-5L dimensions, mobility and anxiety/depression, when using UIRT and MIRT in the MIC dataset. It can be seen that while mobility has distinct ordered categories each of which is the most likely over some of the range of the latent trait covered in both models, anxiety/depression reports shallower slopes over the trait in the UIRT compared to the MIRT, as a result of the lower discrimination.

Table 2 reports the EQ-PRSM UIRT and MIRT models based on the expected a posteriori approach.

Descriptive comparison of PRSM models

Table 5 shows dimension impact for the EQ-PRSM models. Anxiety/depression has the largest impact for the RS EQ VAS weighted, followed by and pain/discomfort and usual activities, while self-care has the lowest impact. For the UIRT models, usual activities and self-care have the largest impact, followed closely by mobility, while and pain/discomfort has the lowest impact. The MIRT model shows a different impact with mobility having the largest impact, followed by anxiety/depression and pain/discomfort and self-care have the mildest impact, For all three models, differences between the three dimensions having the largest impact were relatively small.

Figures 4A-B show the distribution of all modelled EQ-PRSM scores for the two RS and two IRT models, respectively. the EQ VAS weighted algorithm resulted in a smoother modelled summary score distribution, compared to a simple additive approach. The two IRT models appear rather similar, although these distributions do not reveal the relative impact of dimensions on the summary scores on different parts of the scale. Figures 4C shows a comparison of the most promising RS and MIRT models.

Discussion

This proof of principle study demonstrates the usefulness of two different methodological approaches to arrive at PRSM scores for EQ-5D-5L. Based on different theoretical and conceptual backgrounds, RS and IRT approaches both proved that is it feasible to develop PRSM scoring algorithms for EQ-5D-5L, for use in population health and health systems applications. This study was set up to explore innovative ways of deriving PRSM scores for EQ-5D, pushing the boundaries of existing psychometric approaches. In doing so, several steps and assumptions for both RS and IRT approaches could be challenged.

Both approaches have their strengths and weaknesses. The RS approach has a strong theoretical and empirical foundation, and produces scores that are relatively easy to interpret. However, there are known biases that could have affected the models, such as end aversion bias [43]. For this proof of principle study, we applied equal weights across the available population groups for the average RS scores. Considering that a PRSM model would be applied across all conditions, diseases and populations, an alternative approach could be to use prevalence weighting across data from the most prevalent health conditions. A final consideration is that in future research, it could be explored to correct for response heterogeneity when developing PRSM scores using the RS approach.

Note that when developing a patient-reported summary score for non-economic purposes, there is no clear need for country-specific scores since the summary scores are not meant to be used for purposes of resource allocation for a particular country, under the societal perspective. It might be much more sensible to develop patient-or condition-specific summary scores, as evidence presented here shows that scores vary over population groups.

The IRT approach is a firmly established and powerful method for developing scales, although our application in multidimensional health and health-related quality of life instruments differs from most standard applications. We calculated 1PL Rasch and 2PL GRM models. Compared to the 1PL Rasch model, the 2PL UIRT GRM model is theoretically preferable, as it is reasonable to assume that items within health measures are not equally strongly related to the latent trait measured. The 2PL GRM is also theoretically preferable to other 2PL models e.g., generalized partial credit model, as it is easier to interpret. This modelling framework has been already used, due to its flexibility, to examine scale properties, to calibrate items for item banks and score responses for PRO use [32, 44, 45].

From the IRT results it is clear that UIRT ultimately leads to the largest impact on PRSM scores for the first three 'physical' EQ-5D dimensions, which is in line with a study by Feng et al (2019), exploring the internal structure of the EQ-5D [36], who additionally also found that especially anxiety/depression does not fit the unidimensional structure.

There are a few issues with our approach that need to be addressed. Uniform differential item functioning (DIF) may occur when participants with the same score level endorse items differently due to characteristics other than their health, which affects threshold parameters. Examples of these characteristics are language, gender, education, age, health condition, etc. Nonuniform DIF appears in the discrimination parameter and suggests interaction between the underlying measured variable and group membership, which means that the degree to which an item relates to the underlying construct depends on the group being measured. These differences may affect the scores of the IRT models and should be investigated when generating a non-preference based score for the EQ-5D, as recent studies have shown that the EQ-5D, among other instruments, may also be affected by uniform DIF [46].

In this proof of principle study we experimented with using MIRT for the purpose of developing a non-preference based scoring system for the EQ-5D. We have shown that when fitting a multidimensional model, the discrimination parameter of the anxiety/depression improved substantially. This improvement comes at the cost of some additional uncertainty. One kind of such uncertainty is the need of using items from other instruments to estimate the different factors,

which may affect the identification of the latent trait. Such uncertainty may be reduced by increasing the number of items per latent trait, albeit this increases the computational cost. Another uncertainty relates to the use of multiple latent traits. Future research is warranted to investigate this aspect, for example by using second order factor models, or specification of unidimensional models with a more balanced representation between physical and mental health items.

As reported in this study, the presence of a low discrimination parameter has an impact on the range of the scale for the item. This, counterintuitively, impacts the first scoring approach by assigning lower values for that item. By contrast, when using the second approach, items with low discrimination get lower weights as an effect of the use of prior probabilities. While the second approach may be preferable, the results still highlight the need of accounting for the unbalance between mental and physical health dimensions in the EQ-5D if scoring the instrument using IRT.

Apart from the issues discussed above, there are a few additional limitations that need to be addressed. A practical limitation of the RS approach was that data did not allow for all level categories for all dimensions to be used for all population groups, as only categories with 10 or more observations were used. For some dimensions like mobility and self-care, only a few population averages could be used, which could lead to a bias in the results as there were notable differences in average RS scores. We also found a substantial proportion of 'inconsistent' response pairs between the EQ-5D-5L dimensions and corresponding RS scales, where RS scores obviously were reversed. This could be due to the fact that the most severe anchor was placed to the left of the scale, while the first response option for the EQ-5D-5L dimensions, albeit from vertically from top to bottom, is starting with the response option indication no problems. We chose for this RS operationalization because we wanted 0 to indicate the most severe response and 100 the best, similar to the EQ VAS. For future studies a vertical RS approach could accommodate for this apparent phenomenon, or using a face-to-face or computer assisted interviewing mode of administration. There were also differences between the datasets used, in terms of mode of administration, timeframe of data collection, and selection of included countries, that could have led to systematic differences in EQ-5D scores.

Conclusion

This study presents two alternative methods to arrive at patient-reported summary scores for EQ-5D-5L, for use in population health and health systems applications. The methodological approaches can be applied to any scalable multidimensional health instrument. In a follow up study we plan to empirically test the different summary scores for measurement properties.

References

[1] Devlin N, Roudijk B, Ludwig K (eds). Value Sets for EQ-5D-5L: A Compendium, Comparative Review & User Guide. Springer, Dordrecht, 2022.

[2] Devlin N, Parkin D, Janssen B. Methods for Analysing and Reporting EQ-5D Data. Springer: Dordrecht, 2020.

[3] Longworth L, Singh J, Brazier JE. An evaluation of the performance of EQ-5D: a review of reviews of psychometric properties. Paper presented at the 31st EuroQol Scientific Plenary Meeting. 25-26 September 2015, Hotel Birger Jarl: Stockholm, Sweden.

[4] Derrett S, Black J, Herbison GP. Outcome after injury-a systematic literature search of studies using the EQ-5D. J Trauma. 2009 Oct;67(4):883-90. Review.

[5] Dyer MT, Goldsmith KA, Sharples LS, Buxton MJ. A review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes. 2010 Jan 28;8:13.

[6] Ernstsson O, Janssen MF, Heintz E. Collection and use of EQ-5D for follow-up, decision-making, and quality improvement in health care - the case of the Swedish National Quality Registries. J Patient Rep Outcomes. 2020 Sep 16;4(1):78. doi: 10.1186/s41687-020-00231-8

[7] ?

[8] Long D, Haagsma JA, Janssen MF et al. Health-related quality of life and mental well-being of healthy and diseased persons in 8 countries: Does stringency of government response against early COVID-19 matter? SSM Popul Health. 2021 Sep 1;15:100913. doi: 10.1016/j.ssmph.2021.100913

[9] Spronk I, Haagsma JA, Lubetkin EI et al. Health Inequality Analysis in Europe: Exploring the Potential of the EQ-5D as Outcome. Front Public Health. 2021 Nov 4;9:744405. doi: 10.3389/fpubh.2021.744405

[10] Appleby J, Devlin N, Parkin D. (2015) Using Patient Reported Outcomes to Improve Health Care. Wiley Blackwell. ISBN: 978-1-118-94858-3ISBN: 978-1-118-94858-3.

[11] Bansback N, Trenaman L, MacDonald KV, Hawker G, Johnson JA, Stacey D, Marshall DA. An individualized patient-reported outcome measure (PROM) based patient decision aid and surgeon report for patients considering total knee arthroplasty: protocol for a pragmatic randomized controlled trial. BMC Musculoskelet Disord. 2019 Feb 23;20(1):89. doi: 10.1186/s12891-019-2434-2

[12] Pickard AS, Wilke C, Lin HW, Lloyd A. EQ-5D health utilities in studies of cancer: PharmacoEconomics. Pharmacoeconomics. 2007;25(5):365-384.

[13] Janssen MF, Lubetkin E, Sekhobo J, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes. Diabetic Medicine 2011;28: 395-411.

[14] Janssen MF, Birnie E, Bonsel GJ. Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. Qual Life Res. 2008 Apr;17(3):463-73.

[15] Hout, B van, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health 2012 Jul-Aug;15(5):708-15.

[16] Janssen MF, Pickard AS et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L: a multicountry study. Qual Life Res. 2013; 22(7): 1717–1727.

[17] Voormolen DC, Polinder S, von Steinbuechel N et al. Health-related quality of life after traumatic brain injury: deriving value sets for the QOLIBRI-OS for Italy, The Netherlands and The United Kingdom. Qual Life Res. 2020 Nov;29(11):3095-3107. doi: 10.1007/s11136-020-02583-6

[18] Richardson J, Iezzi A, Maxwell A. Cross-national comparison of twelve quality of life instruments. Research papers 78. 80-83. 85. MIC Report 2. Centre for Health Economics. Monash University. 2012.

[19] Huskisson EC. Measurement of pain. Lancet 1974, ii, 1127-31.

[20] Aitken RCB. A growing edge of measurement of feelings. Proceedings of the Royal Society of Medicine, 1969, 62, 989-92.

[21] Scott PJ and Huskisson EC. Measurement of functional capacity with visual analouge scales. Rheumatology and rehabilitation. 1979, 16, 257-9.

[22] Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L Better Than EQ-5D-3L? A Head-to-Head Comparison of Descriptive Systems and Value Sets from Seven Countries. Pharmacoeconomics. **2018b** Feb 22. doi: 10.1007/s40273-018-0623-8

[23] Richardson J, Sinha K, Iezzi A, Khan MA. Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. Qual Life Res. 2014 Oct;23(8):2395-404.

[24] Sintonen, H. (1995). The 15D-measure of health-related quality of life. II. Feasibility, reliability and validity of its valuation system. Natl Cent Health Program Eval Work Pap 42 Melbourne.

[25] van der Weijden T, Légaré F, Boivin A, Burgers JS, van Veenendaal H, Stiggelbout AM, Faber M, Elwyn G. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. Implement Sci. 2010 Feb 2;5:10. doi: 10.1186/1748-5908-5-10.
[26] Armstrong MJ et al. Value Assessment at the Point of Care: Incorporating Patient Values throughout Care Delivery and a Draft Taxonomy of Patient Values. Value Health. 2017 Feb;20(2):292-295.
[27] Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. Soc Sci Med. 2016 Sep;165:66-74
[28] Burström K, Sun S, Gerdtham U, Henriksson M, Johannesson M, Levin L, et al. Swedish experiencebased value sets for EQ-5D health states. Quality of life research. 2014;23(2):431-42.
[29] Leidl R, Reitmeir P. An Experience-Based Value Set for the EQ-5D-5L in Germany. Value Health. 2017 Sep;20(8):1150-1156.
[30] Sameijma, F. Graded response model. In: van der Linden W. & Hambleton R. (eds.) Handbook of modern item response theory. (pp. 85–100). Berlin: Springer.
[31] Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Nielsen and Lydiche.
[32] Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, et al. (2007) Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 45: S22–S31. Available: http://www.ncbi.nlm.nih.gov/pubmed/17443115. Accessed 2014 July 12. PMID: 17443115
[33] Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4:272.
[34] Bentler PM, Dudgeon P. Covariance structure analysis: statistical practise, theory and directions. Annu Rev Psychol 1996;47:541–70.
[35] Feng YS, Jiang R, Pickard AS et al. Combining EQ-5D-5L items into a level summary score: demonstrating feasibility using non-parametric item response theory using an international dataset. Qual Life Res. 2022 Jan;31(1):11-23. doi: 10.1007/s11136-021-02922-1
[36] Feng YS, Jiang R, Kohlmann T et al. Exploring the Internal Structure of the EQ-5D Using Non-Preference-Based Methods. Value Health. 2019 May;22(5):527-536. doi: 10.1016/j.jval.2019.02.006
[37] Finch AP, Brazier JE, Mukuria et al. An Exploratory Study on Using Principal-Component Analysis and Confirmatory Factor Analysis to Identify Bolt-On Dimensions: The EQ-5D Case Study. Value Health. 2017 Dec;20(10):1362-1375. doi: 10.1016/j.jval.2017.06.002
[38] McDonald, RP 1981. The dimensionality of tests and items. British Journal of mathematical and statistical Psychology, 34, 100-117.
[39] Bock RD, Aitkin M Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika. 1981; 46(4):443–459.
[40] Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982; 6(4):431–44
[41] Chapman R. Expected a posteriori scoring in PROMIS. Journal of Patient Reported Outcomes. 2022; 6:59.
[42] Vanden Bos GR. APA dictionary of psychology, 1st edn. American Psychological Association, Washington. 2007.
[43] Hollingworth H.L. The central tendency of judgment. The Journal of Philosophy, Psychology and Scientific Methods. 1910;7(17),461–469.
[44] Mukuria et al. Deriving a Preference-Based Measure for Myelofibrosis from the EORTC QLQ-C30 and the MF-SAF. 2015.
[45] Dewitt et al., Estimation of a Preference-Based Summary Score for the Patient-Reported OutcomesMeasurement Information System: The PROMIS Preference (PROPr) Scoring System. 2018
[46] Penton H, Dayson C, Hulme C et al. An Investigation of Age-Related Differential Item Functioning in the EQ-5D-5L Using Item Response Theory and Logistic Regression. Value Health. 2022 Apr 26;S1098-3015(22)00151-6. doi: 10.1016/j.jval.2022.03.009

Table 1. Sample characteristics of modeling data and datasets used by approach (total N=22,050)

Population	Country	N	% female	Mean age (years)	Mean EQ VAS (SD)	RS approach*	UIRT approach	MUIRT approach
Crosswalk dataset								
Asthma/COPD	England, Scotland	342	52	67	58 (21)	√	√	√
Cardiovascular disease	England, Scotland	251	46	67	61 (21)	√	√	√
Depression	England	250	56	42	62 (21)	√	√	√
Diabetes	Denmark, England	276	48	52	74 (20)	√	√	√
Liver disease	Italy	619	36	57	70 (21)		√	√
Personality disorder	Netherlands	380	67	32	60 (18)	√	√	√
RA/Arthritis	Denmark, England, Scotland	369	52	61	63 (21)	√	√	√
Stroke	England, Poland	596	47	68	53 (26)	√	√	√
Other**	Denmark, England, Netherlands, Scotland	330	46	45	64 (24)	√	√	√
Students	Poland	443	79	22	79 (16)		√	√
QOLIBRI dataset								
General population	Italy, Netherlands, UK	10,172	46	45	75 (20)	√		
MIC dataset								
General population	Australia, Canada, Germany, UK, US	8,022	52	52	67 (22)			√

VAS visual analogue scale, RS rating scale, IRT item response theory, UIRT unidimensional item response theory, MIRT multidimensional item response theory COPD chronic obstructive pulmonary disease, RA rheumatoid arthritis, MIC multi instrument comparison, UK United Kingdom, US United States

*Data with RS scales available

**Back pain, ADHD, kidney dialysis, multiple sclerosis, orthopaedic accident, Parkinson's

Table 2. EQ-PRSM models for RS and IRT approaches (UIRT based on the crosswalk data, MIRT on the MIC data)

EQ-PRSM models		**RS approach**		**IRT approach***	
		Equal weights	EQ VAS weights	UIRT	MIRT
Mobility	Slight	5.07	4.51	10.53	12.10
	Moderate	9.75	8.67	14.34	15.08
	Severe	15.28	13.58	18.34	19.86
	Unable to	20.00	17.78	25.90	25.67
Self-care	Slight	6.04	2.62	7.71	2.32
	Moderate	11.54	5.01	12.05	3.48
	Severe	16.55	7.18	17.59	5.65
	Unable to	20.00	8.68	26.47	9.38
Usual activities	Slight	4.82	5.67	13.63	10.18
	Moderate	10.04	11.81	18.58	12.62
	Severe	15.79	18.57	22.15	15.04
	Unable to	20.00	23.52	26.94	16.76
Pain/discomfort	Slight	4.10	4.89	7.77	7.02
	Moderate	9.31	11.09	8.97	12.92
	Severe	15.37	18.30	10.73	17.53
	Extreme	20.00	23.82	13.55	23.42
Anxiety/depression	Slight	5.46	7.15	3.10	7.32
	Moderate	10.77	14.11	4.26	13.20
	Severe	16.21	21.24	5.35	18.40
	Extreme	20.00	26.21	7.14	24.77

PRSM patient-reported summary score, RS rating scale, VAS visual analogue scale, IRT item response theory, UIRT unidimensional item response theory, MIRT multidimensional item response theory

*Based on the expected a posteriori approach

Table 3. Difficulty and discrimination parameters of UIRT GRM (crosswalk and MIC data) and MIRT GRM (MIC data), range and rank

		Difficulty						
	Dimensions	Slight (b1)	Moderate (b2)	Severe (b3)	Extreme (b4)	**Range**	**Discrimination (a)**	**Rank**
UIRT				Crosswalk dataset				
	Mobility	0.183	-0.380	-1.036	-1.914	2.097	3.635	2
	Selfcare	-0.398	-0.880	-1.390	-1.805	1.407	4.873	1
	Usual activities	0.465	-0.243	-0.911	-1.583	2.048	3.480	3
	Pain	0.774	-0.263	-1.381	-2.647	3.421	1.808	4
	Anxiety	0.650	-0.856	-2.461	-4.290	4.940	0.885	5
UIRT				MIC dataset				
	Mobility	-0.461	-1.149	-1.882	-2.924	2.463	3.829	2
	Selfcare	-1.328	-1.974	-2.727	-3.543	2.215	3.145	3
	Usual activities	-0.399	-1.192	-1.934	-2.653	2.254	4.506	1
	Pain	0.677	-0.621	-1.578	-2.737	3.414	2.452	4
	Anxiety	0.012	-1.649	-3.130	-4.556	4.568	0.942	5
MUIRT				MIC dataset				
	Mobility	-0.448	-1.115	-1.781	-2.697	2.249	5.716	3
	Selfcare	-1.347	-1.982	-2.727	-3.546	2.199	3.038	5
	Usual activities	-0.423	-1.246	-2.001	-2.752	2.329	3.549	4
	Pain	0.543	-0.515	-1.334	-2.226	2.769	8.112	1
	Anxiety	0.010	-0.840	-1.571	-2.179	2.189	6.336	2

UIRT unidimensional item response theory, MIRT multidimensional item response theory, MIC multi instrument comparison

Table 4. UIRT and MIRT models using most representative level of the theta approach

		Method 1*				Method 2**			
	Dimension	Slight	Moderate	Severe	Extreme	Slight	Moderate	Severe	Extreme
UIRT		Crosswalk dataset				Crosswalk dataset			
	Mobility	-0.099	-0.709	-1.469	-2.179	-0.519	-0.707	-0.904	-1.277
	Selfcare	-0.639	-1.129	-1.599	-1.919	-0.380	-0.594	-0.867	-1.305
	Usual activities	0.110	-0.579	-1.249	-1.719	-0.672	-0.916	-1.092	-1.328
	Pain	0.250	-0.819	-2.009	-2.919	-0.383	-0.442	-0.529	-0.668
	Anxiety	-0.099	-1.659	-3.379	-5.131	-0.153	-0.210	-0.264	-0.352
UIRT		MIC dataset				MIC dataset			
	Mobility	-0.809	-1.519	-2.399	-3.209	-0.429	-0.565	-0.791	-1.013
	Selfcare	-1.649	-2.349	-3.049	-3.859	-0.197	-0.283	-0.466	-0.703
	Usual activities	-0.799	-1.559	-2.289	-2.889	-0.540	-0.779	-1.036	-1.270
	Pain	0.150	-1.099	-2.159	-3.149	-0.642	-0.827	-0.912	-1.023
	Anxiety	-0.809	-2.389	-3.839	-5.299	-0.170	-0.220	-0.252	-0.264
MUIRT		MIC dataset				MIC dataset			
	Mobility	-0.779	-1.449	-2.239	-2.719	-0.724	-0.902	-1.188	-1.536
	Selfcare	-1.659	-2.349	-3.139	-3.699	-0.139	-0.208	-0.338	-0.561
	Usual activities	-0.829	-1.619	-2.379	-2.869	-0.609	-0.755	-0.900	-1.003
	Pain	0.010	-0.919	-1.779	-2.229	-0.420	-0.773	-1.049	-1.401
	Anxiety	-0.409	-1.209	-1.869	-2.219	-0.438	-0.790	-1.101	-1.482

UIRT unidimensional item response theory, MIRT multidimensional item response theory, MIC multi instrument comparison

*Theta's reported

**Expected a posteriori approach: regression (B coefficients) reported, all statistically significant at $p<0.01$

Table 5. Dimension impact* for EQ-PRSM models (UIRT based on the crosswalk data, MIRT on the MIC data)

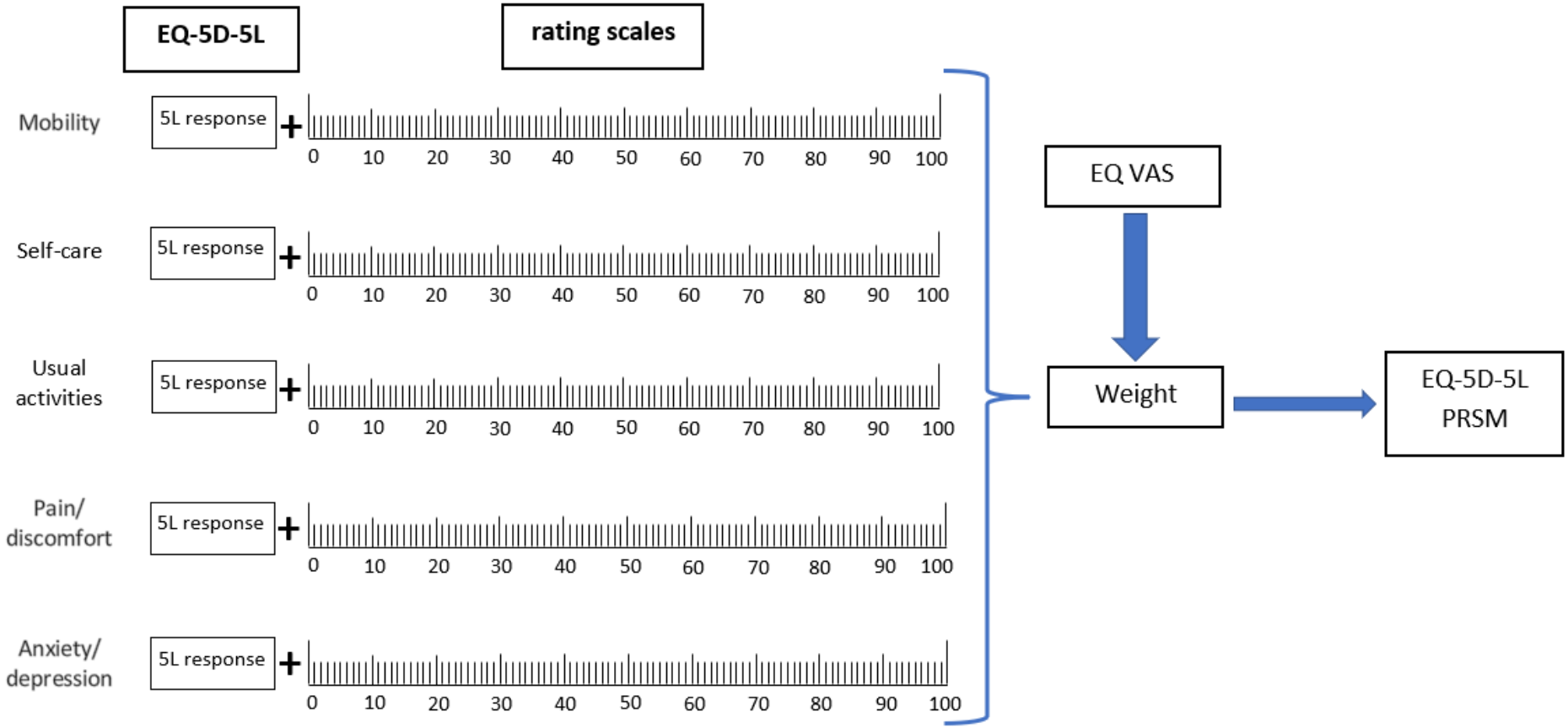
EQ-PRSM models	RS approach		IRT approach	
	Equal weights**	EQ VAS weights	UIRT	MIRT
	NA	Anxiety/depression	Usual activities	Mobility
	NA	Pain/discomfort	Self-care	Anxiety/depression
	NA	Usual activities	Mobility	Pain/discomfort
	NA	Mobility	Anxiety/depression	Usual activities
	NA	Self-care	Pain/discomfort	Self-care

PRSM patient-reported summary score, RS rating scale, VAS visual analogue scale, IRT item response theory, UIRT unidimensional item response theory, MIRT multidimensional item response theory, NA not applicable

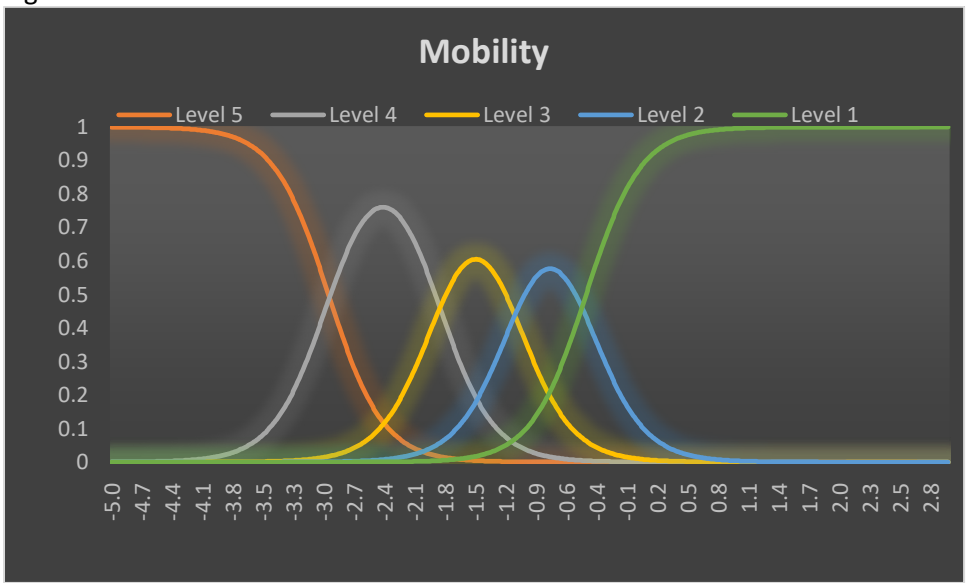
*Impact is judged by the size of the coefficient for level 5 in each dimension

**By definition all coefficients are 20 for the equal weights model

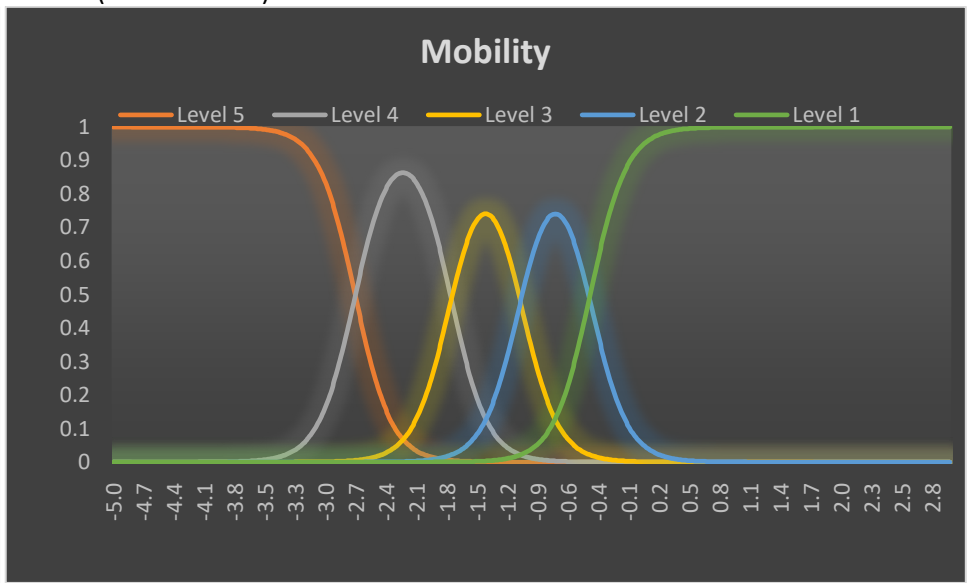
Figure 1. Schematic overview of RS approach



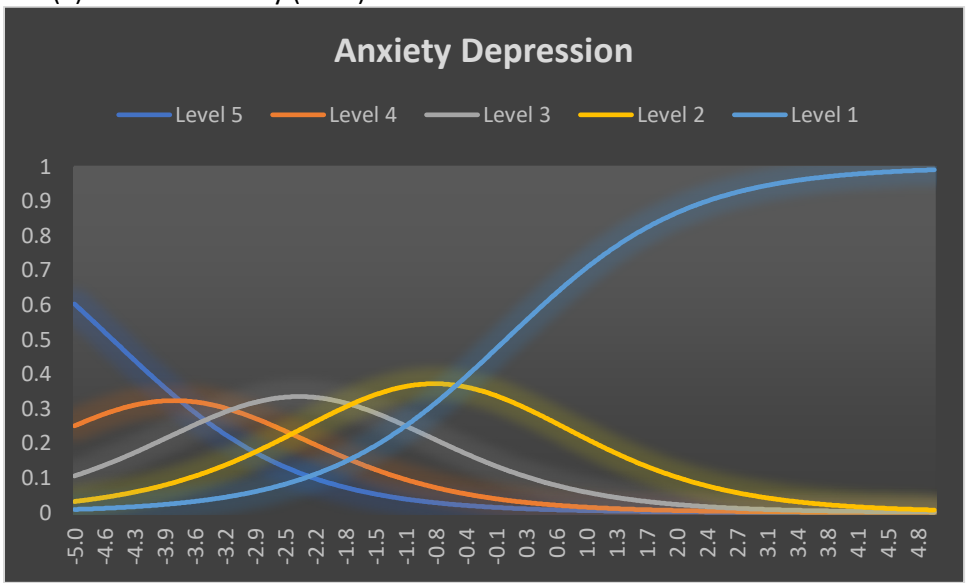
Figures 2A-D. Item characteristics curves for the UIRT GRM model and MIRT GRM model (MIC datasets)



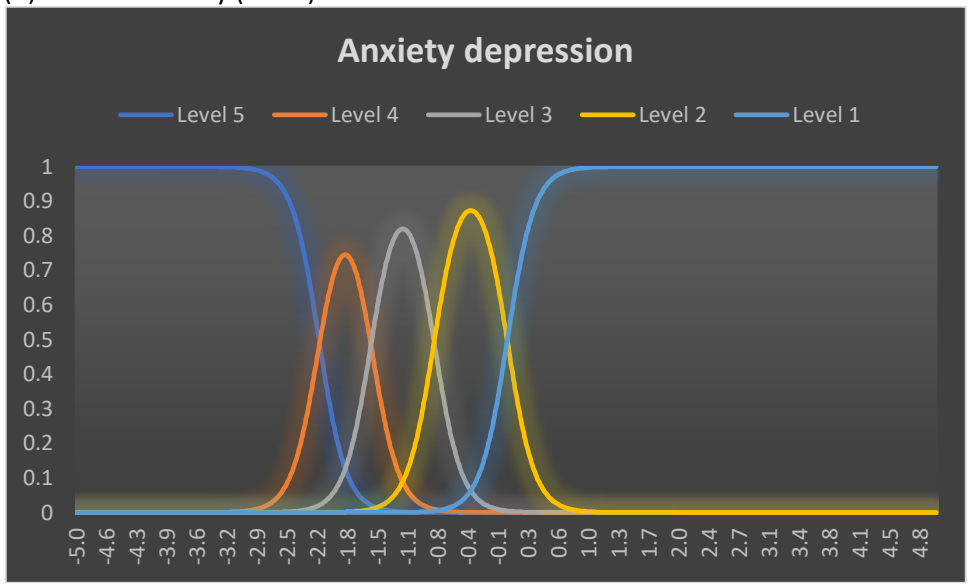
(a) ICC for Mobility (UIRT)



(b) ICC for Mobility (MIRT)

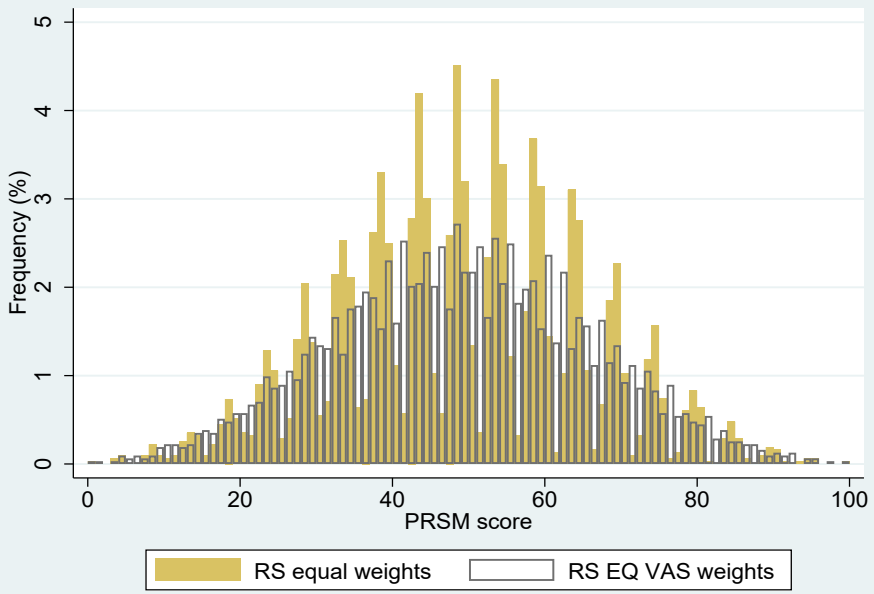


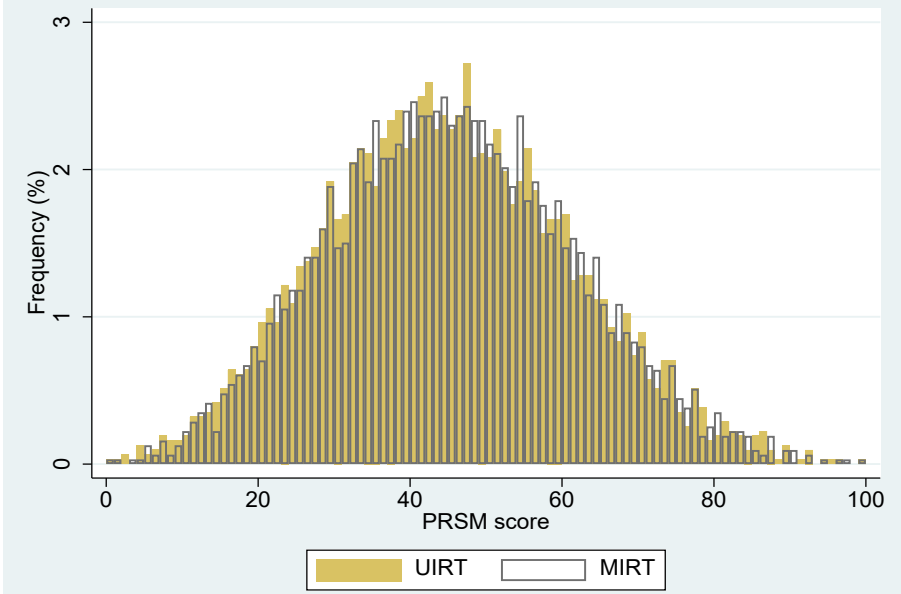
(c) ICC for Anxiety / Depression (UIRT)

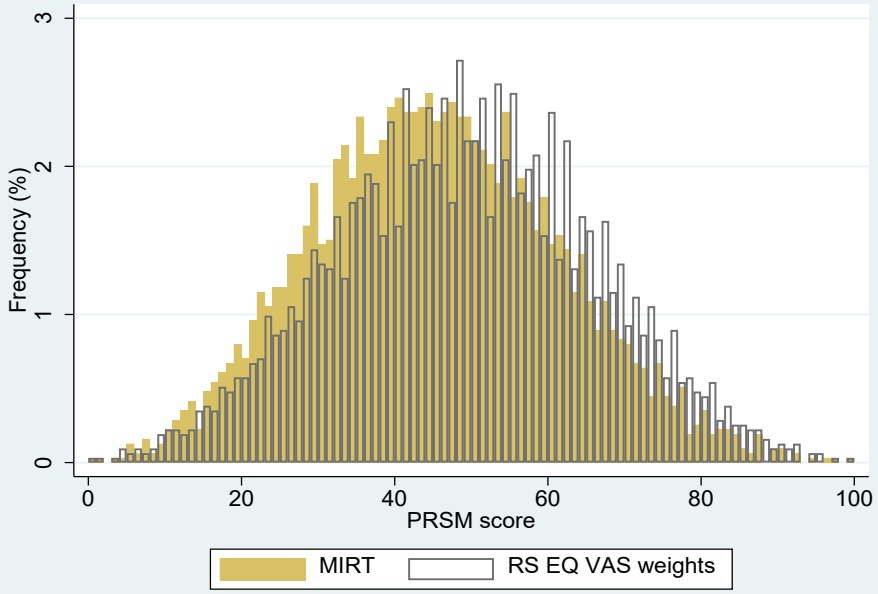


(d) ICC for Anxiety / Depression (MIRT)

Figures 3A-B. Histograms of all possible EQ-PRSM scores for two RS and two IRT models, and preferred RS vs IRT models and RS vs LSS







Appendix 1. Level distributions and mean RS scores for crosswalk, QOLIBRI and MIC datasets*

Dimension	Population	Asthma/COPD			Cardiovascular disease			Depression			Diabetes			Liver disease			Personality disorder			RA/Arthritis			Stroke			Other**			Students			QOLIBRI			MIC		
				Mean			Mean			Mean			Mean			Mean			Mean			Mean			Mean			Mean			Mean			Mean			Mean	
		N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS	N	%	RS
Mobility	No	72	21	90	56	22	92	154	62	95	177	64	97	453	73	NA	318	84	97	83	22	91	118	20	96	116	35	95	428	97	NA	7,670	75	93	1400	43	NA
	Slight	80	23	71	60	24	69	54	22	72	52	19	78	99	16	NA	39	10	73	115	31	73	116	19	70	83	25	67	12	3	NA	1,437	14	70	664	21	NA
	Moderate	94	27	54	74	29	53	24	10	52	26	9	52	50	8	NA	21	6	56	101	27	54	155	26	55	68	21	47	2	0	NA	714	7	51	610	19	NA
	Severe	90	26	32	56	22	30	17	7	21	21	8	34	16	3	NA	1	0	30	67	18	30	111	19	31	41	12	34	1	0	NA	288	3	31	417	13	NA
	Unable to	6	2	3	5	2	19	1	0	0	0	0		1	0	NA	1	0	0	3	1	39	96	16	3	22	7	30	0	0	NA	63	1	31	140	4	NA
Self-care	No	192	56	96	136	54	96	204	82	93	229	83	98	554	89	NA	355	93	97	223	60	92	188	32	90	186	56	91	442	100	NA	9,077	89	93	2091	65	NA
	Slight	70	20	71	61	24	70	21	8	68	36	13	71	41	7	NA	21	6	74	84	23	72	121	20	67	62	19	64	0	0	NA	608	6	64	499	16	NA
	Moderate	52	15	48	35	14	51	21	8	45	8	3	54	19	3	NA	3	1	62	43	12	46	114	19	52	44	13	52	1	0	NA	351	3	48	336	11	NA
	Severe	19	6	30	12	5	24	4	2	33	3	1	20	4	1	NA	1	0	30	17	5	34	57	10	30	30	9	37	0	0	NA	95	1	35	150	4	NA
	Unable to	9	3	13	7	3	19	0	0		0	0		1	0	NA	0	0		2	1	0	116	19	5	8	2	47	0	0	NA	41	0	44	146	4	NA
Usual activities	No	76	22	93	64	25	93	113	45	95	160	58	98	422	68	NA	98	26	94	81	22	94	106	18	95	94	28	91	376	85	NA	7,466	73	92	1067	33	NA
	Slight	91	27	75	57	23	73	72	29	73	69	25	76	101	16	NA	85	22	73	131	36	73	127	21	71	80	24	70	48	11	NA	1,571	15	71	788	25	NA
	Moderate	87	25	51	67	27	55	37	15	51	28	10	52	69	11	NA	119	31	58	94	25	51	138	23	50	69	21	53	15	3	NA	767	8	50	688	21	NA
	Severe	66	19	31	42	17	33	25	10	26	13	5	32	22	4	NA	68	18	42	46	12	28	94	16	29	47	14	37	3	1	NA	279	3	33	417	13	NA
	Unable to	22	6	12	21	8	13	3	1	14	6	2	7	5	1	NA	10	3	19	17	5	15	131	22	6	40	12	25	1	0	NA	89	1	27	254	8	NA
Pain/discomfort	No	76	22	89	64	25	91	82	33	92	115	42	98	361	58	NA	137	36	92	26	7	91	115	19	89	87	26	89	268	60	NA	5,002	49	89	937	29	NA
	Slight	88	26	74	71	28	75	88	35	76	92	33	74	146	24	NA	132	35	73	123	33	75	146	24	69	85	26	69	143	32	NA	3,126	31	72	931	29	NA
	Moderate	105	31	53	61	24	48	48	19	56	41	15	64	91	15	NA	83	22	57	135	37	57	209	35	52	103	31	53	29	7	NA	1,383	14	54	851	26	NA
	Severe	60	18	27	45	18	37	24	10	26	23	8	37	19	3	NA	26	7	35	73	20	35	100	17	33	41	12	37	3	1	NA	494	5	38	406	13	NA
	Extreme	13	4	17	10	4	14	8	3	5	5	2	17	2	0	NA	2	1	50	12	3	25	26	4	7	14	4	22	0	0	NA	167	2	31	92	3	NA
Anxiety/depression	No	163	48	93	110	44	94	33	13	89	172	62	97	341	55	NA	51	13	93	190	51	91	121	20	90	157	48	89	190	43	NA	5,866	58	88	1226	38	NA
	Slight	81	24	69	70	28	70	89	36	70	71	26	78	162	26	NA	82	22	72	100	27	66	209	35	63	74	22	66	173	39	NA	2,414	24	66	895	28	NA
	Moderate	74	22	46	51	20	50	80	32	53	25	9	55	92	15	NA	119	31	53	54	15	48	165	28	50	58	18	46	55	12	NA	1,218	12	52	687	21	NA
	Severe	20	6	27	14	6	17	32	13	30	7	3	39	19	3	NA	103	27	36	18	5	26	79	13	36	26	8	52	21	5	NA	375	4	41	313	10	NA
	Extreme	4	1	9	6	2	17	16	6	37	1	0	15	5	1	NA	25	7	12	7	2	17	22	4	6	15	4	24	4	1	NA	299	4	38	99	3	NA

COPD chronic obstructive pulmonary disease, RA rheumatoid arthritis, MIC multi instrument comparison, NA not available

*RS means scores in grey are not used in the analysis as these are based on <10 observations

Appendix 2. EQ VAS regression weights (RS approach) by population (crosswalk and QOLIBRI)

Population	Asthma/COPD		Cardiovascular disease		Depression		Diabetes		Personality disorder		RA/Arthritis		Stroke		Other**		QOLIBRI		Average	
Dimension	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank	Coeff.	Rank
Mobility	3.36	3	3.03	4	4.35	2	2.94	3	1.79	4	3.45	5	4.03	2	4.34	4	3.01	4	3.37	4
Self-care	2.04	5	3.80	3	1.29	5	1.05	5	0.43	5	3.87	3	1.21	5	-0.37	5	1.46	5	1.64	5
Usual activities	4.70	1	1.63	5	4.09	3	2.36	4	4.55	1	4.31	2	7.79	1	5.76	1	4.88	3	4.45	3
Pain/discomfort	2.91	4	4.45	2	3.00	4	7.99	1	4.29	2	5.20	1	2.75	4	5.07	2	4.91	1	4.51	2
Anxiety/depression	4.27	2	5.99	1	8.49	1	5.65	2	4.06	3	3.79	4	3.05	3	4.45	3	4.90	2	4.96	1

RS rating scale, VAS visual analogue scale, COPD chronic obstructive pulmonary disease, RA rheumatoid arthritis