

Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems (CORE-OM)

Ifigenia Mavranzouli · John E. Brazier · Tracey A. Young · Michael Barkham

Accepted: 29 September 2010 / Published online: 23 October 2010
© Springer Science+Business Media B.V. 2010

Abstract

Purpose To describe a new approach for deriving a preference-based index from a condition specific measure that uses Rasch analysis to develop health states.

Methods The CORE-OM is a 34-item instrument monitoring clinical outcomes of people with common mental health problems. The CORE-OM is characterised by high correlation across its domains. Rasch analysis was used to reduce the number of items and response levels in order to produce a unidimensional measure and to generate a credible set of health states corresponding to different levels of symptom severity using the Rasch item threshold map.

Results The proposed methodology resulted in the development of CORE-6D, a 2-dimensional health state descriptive system consisting of a unidimensional 5-item emotional component (derived from Rasch analysis) and a physical symptom item. Inspection of the Rasch item threshold map of the emotional component helped identify

a set of 11 plausible health states, which, combined with 3 physical symptom item levels, form 33 plausible health states that can be used for the valuation of the instrument, resulting in the development of a preference-based index. **Conclusions** This is a useful new approach to develop preference-based measures from existing instruments with high correlations across domains. The CORE-6D preference-based index will enable calculation of Quality-Adjusted Life Years in people with common mental health problems.

Keywords · Rasch analysis · Health-related quality of life · Condition-specific measure of health · Preference-based measures of health · Health states · CORE-6D · CORE-OM · Mental health · Quality-Adjusted Life Year

Abbreviations

CORE-OM	Clinical Outcomes in Routine Evaluation-Outcome Measure
CSM	Condition-specific measure
DIF	Differential item functioning
HRQoL	Health-related quality of life
NHS	National Health Service
PBM	Preference-based measure
SRM	Standardised response mean
QALY	Quality-Adjusted Life Year

I. Mavranzouli · J. E. Brazier · T. A. Young
Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Sheffield S1 4DA, UK

I. Mavranzouli (✉)
National Collaborating Centre for Mental Health,
Centre for Outcomes Research and Effectiveness (CORE),
Research Department of Clinical, Educational & Health
Psychology, University College London, Gower Street,
London WC1E 6BT, UK
e-mail: i.mavranzouli@ucl.ac.uk

T. A. Young
Yorkshire and Humber Research Design Service,
University of Sheffield, Sheffield S1 4DA, UK

M. Barkham
Centre for Psychological Services Research,
University of Sheffield, Sheffield S10 2TN, UK

Introduction

Calculation of Quality-Adjusted Life Years (QALYs) in cost-utility analysis requires description and subsequent valuation of health states characterising a disease area [1]. Generic preference-based measures (PBMs), such as

EQ-5D [2], SF-6D [3], and HUI3 [4], are widely used for this purpose. These instruments consist of a general health state descriptive system and an algorithm converting each health state into a utility value.

Generic PBMs may be inappropriate or insensitive in capturing Health-Related Quality of Life (HRQoL) in some medical conditions [5]. On the other hand, the majority of available condition-specific measures (CSMs) are not preference-based. One solution to this problem has been the “mapping” from CSMs directly onto generic PBMs (e.g. [6, 7]); however, this process may result in limited performance in terms of model fit and ability to predict values where the overlap between the generic measure and the CSM is poor [8, 9]. For this reason, there has been an increased interest in the development of PBMs directly from existing CSMs.

CSMs normally consist of a large number of items capturing multiple dimensions of health. Inclusion of all items in a PBM would often result in the description of a massive number of potential health states that would be impractical to use and complicated to value. The main approach of dealing with this situation is to develop health state classifications by selecting 1–2 items from each dimension represented in a CSM, thus defining a concise set of health states. This approach was first applied to the generic SF-36 in the development of the SF-6D preference-based index [3] and has since been used at the development of PBMs from a number of CSMs [10–13]. Factor analysis can be used in such cases to assess the dimensional structure of a measure, explore potential correlations between dimensions, and suggest appropriate reductions in dimensions [14]. Items may be selected based on classical psychometric criteria, such as internal consistency and responsiveness to change. Rasch analysis has also been used at the development of health state classifications from existing CSMs, in order to select items within dimension and reduce item response levels [15, 16].

Ideally, health state classifications should have a multi-dimensional structure with little or no correlation between items. This requirement results from the demands of the valuation stage, where a sample of states is selected for valuation since it is not practical to value all states. For instruments like EQ-5D and SF-6D that employ statistical inference, statistical designs such as orthogonal arrays and balanced designs are used to estimate additive models in order to predict the values for all potential health states. For the HUI3 that uses multi-attribute utility theory, ‘corner’ states must be valued where one dimension is at the worst level and all others are at the best level. A major problem arises when items in a health state classification tap the same or highly correlated dimensions and therefore cannot be treated independently, as separate statements. In such cases, some of the health states may include combinations of statements that are not plausible (e.g. ‘I feel happy most

of the time’ and ‘I often feel like crying’). This problem is most likely to arise in the case of CSMs with high correlation between dimensions.

One approach for developing plausible health states from measures characterised by high correlation across their items has been described by Sugar et al. [17]. The authors conducted cluster analysis in order to identify distinct groups of patients with depression based on their mental and physical health composite scores on SF-12. The resulting patient groups corresponded to 6 distinct health state descriptions for depression that were clinically meaningful since they were derived from actual cases observed in clinical practice. This approach can therefore be employed for the development of PBMs from CSMs with few and highly correlated dimensions, where conventional statistical approaches for generating health states (such as orthogonal arrays) are not appropriate.

In this paper, we propose an alternative approach for constructing plausible health state descriptions from CSMs with high correlation between their dimensions, using Rasch analysis. Rasch analysis has already been used in order to select appropriate items and response levels from existing multidimensional CSMs [15, 16]. Here, we take advantage of another property of Rasch models relevant to our context, that is, the ability of Rasch analysis to assign respondents to different points of severity along the latent variable, based on their responses, and to subsequently generate groups of respondents of different symptom severity [18]. We have used this attribute of Rasch models in order to develop plausible health states from the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM).

Methods

The Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM)

The CORE-OM is an instrument measuring common mental health problems that has been developed to evaluate the effectiveness of psychological therapies across multi-disciplinary services in the United Kingdom [19]. It consists of 34 items, each with 5 levels of response: ‘not at all’, ‘only occasionally’, ‘sometimes’, ‘often’, and ‘most or all the time’. The items tap 4 conceptual domains: ‘subjective well-being’ (4 items), ‘problems’ (4 items on depression, 4 items on anxiety, 2 items on physical symptoms, and 2 items on trauma), ‘functioning’ (4 items on general functioning, 4 items on close relationships, and 4 items on social relationships), and ‘risk’ (4 items on risk-to-self and 2 items on risk-to-others). Eight of the items are positively worded. The dimensional structure of CORE-OM is presented in Table 1.

Table 1 The dimensional structure of the CORE-OM

Domain	Item No	Item
Subjective well-being	4	I have felt ok about myself
	14	I have felt like crying
	17	I have felt overwhelmed by my problems
	31	I have felt optimistic about my future
Symptoms—anxiety	2	I have felt tense, anxious or nervous
	11	Tension/anxiety have prevented me doing important things
	15	I have felt panic or terror
	20	My problems have been impossible to put to one side
Symptoms—depression	5	I have felt totally lacking in energy and enthusiasm
	23	I have felt despairing or hopeless
	27	I have felt unhappy
	30	I have thought I am to blame for my problems & difficulties
Symptoms—physical	8	I have been troubled by aches, pains, physical problems
	18	I have had difficulty of getting to sleep or staying asleep
Symptoms—trauma	13	I have been disturbed by unwanted thoughts and feelings
	28	Unwanted images or memories have been distressing me
Functioning—general	7	I have felt able to cope when things go wrong
	12	I have been happy with the things I have done
	21	I have been able to do most things I needed to
	32	I have achieved the things I wanted to
Functioning—close relationships	1	I have felt terribly alone and isolated
	3	I have felt I have somebody to turn to for support when needed
	19	I have felt warmth or affection for someone
	26	I have thought I have no friends
Functioning—social relationships	10	Talking to people has felt too much for me
	25	I have felt criticised by other people
	29	I have been irritable when with other people
	33	I have felt humiliated or shamed by other people
Risk/harm to self	9	I have thought of hurting myself
	16	I made plans to end my life
	24	I have thought it would be better if I were dead
	34	I have hurt myself physically or taken risks with my health
Risk/harm to others	6	I have been physically violent to others
	22	I have threatened or intimidated another person

The CORE-OM comprises a valid, reliable, and acceptable effectiveness measure across a wide range of practice settings offering psychological therapies [20, 21]. It has been routinely used to evaluate psychological therapies and counselling services in primary and secondary settings in the United Kingdom [19, 22] and is a widely used patient-based tool for measuring mental health outcomes in the British National Health Service (NHS) [23, 24]. Based on these characteristics and given the scepticism about use of generic PBMs in mental health and the arguments favouring the development of a mental health-specific PBM [25–27], CORE-OM was selected as the basis for constructing a PBM specific to common mental health problems.

With 34 items having 5 levels each, CORE-OM may form a practically unmanageable number of 5^{34} health states. Previously undertaken exploratory factor analysis revealed that the 34 items loaded on 3 components, one including mainly the negatively worded items, one made up of the positively worded items, and a risk items component [20]. The same study examined the correlation across the instrument domains and demonstrated that the domains of ‘subjective well-being’, ‘problems’, and ‘functioning’ were highly correlated with each other (in pairwise examinations of the 3 domains the Spearman’s ρ value exceeded 0.70 in both clinical and non-clinical populations); the ‘risk’ items also showed high though somewhat lower correlation with the non-risk items

(Spearman's ρ value = 0.64 in a clinical sample; 0.44 in a non-clinical sample) [20]. Thus generating states using standard statistical design from the health state classification would not be appropriate in this case, as it would most likely result in implausible health states. For this reason, a new method using Rasch analysis was applied, aiming at the construction of credible health states from CORE-OM.

The CORE-OM dataset used in Rasch analysis

Data analysed in this study were derived from a database containing patient information from 33 NHS primary care services in the United Kingdom. Data included CORE-OM scores, as well as patients' age, gender, and ethnicity. Details on the database and the data collection procedures are available in Evans et al. [22]. A sample of 1,500 primary care clients formed the dataset for the work presented in this paper [N1500]. Conventional psychometric tests were conducted on the whole dataset [N1500]. A random sample of 400 respondents [N400a] out of the whole dataset [N1500] was used in Rasch analysis. Use of a smaller sample size for Rasch analysis was dictated by evidence that some Rasch fit statistics for polytomous scales (like CORE-OM) are highly dependent on sample size, which translates into a higher possibility for type I errors with increased sample size [28]. The findings of Rasch analysis on [N400a] were validated on another random sample of 400 respondents [N400b] out of the whole dataset [N1500].

Use of Rasch analysis to select items and identify plausible health states amenable to valuation

The Rasch model is underpinned by the principles of unidimensionality and local independence of items [29]. Rasch analysis can therefore be used at the development of unidimensional PBMs derived from existing CSMs with no clear multidimensional structure. The objective of this study was to use Rasch analysis in order to construct a health state descriptive system from CORE-OM, amenable to valuation. This process, which resulted in the development of a measure able to describe health states for common mental health problems across 6 conceptual domains (named 'CORE-6D'), followed 4 major steps illustrated in the flow diagram shown in Fig. 1.

Step 1 Use of Rasch analysis, conventional psychometric tests, and expert opinion in order to exclude items and develop a unidimensional scale

Rasch analysis was undertaken in order to identify and subsequently exclude CORE-OM items not fitting the Rasch model. Conventional psychometric tests provided additional indication of an item's suitability for inclusion in a PBM and were considered at deciding which item to

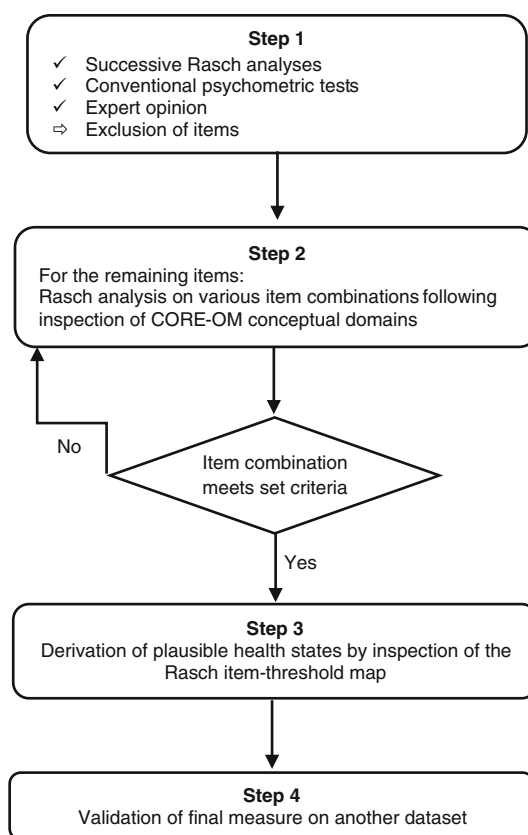


Fig. 1 Flow diagram of the process of developing a health state descriptive system and plausible health states using Rasch analysis as a primary tool

exclude first, when two or more items met exclusion criteria in Rasch analysis. The criteria used to exclude items from Rasch analysis and judge their classical psychometric properties have been described and justified in previous related studies [15, 16]. In summary, the following criteria were considered at the development of the new instrument:

Rasch analysis criteria for item exclusion

- **Item level ordering:** item-threshold maps were inspected to investigate whether respondents were able to distinguish between adjacent response levels. When items had disordered thresholds (i.e. when an item score was likely to decrease as respondent's symptom severity increased), then visual inspection of respective category probability curves determined which adjacent responses to merge. If the only way to order an item's thresholds was by merging adjacent responses that were not clinically meaningful, then this item was excluded. For example, a new response level that was formed by merging response levels 'sometimes' and 'often' was deemed to be not clinically meaningful in any of the items; in addition, response levels 'not at all' and 'only

occasionally' could not be merged in the case of 'severe' items, such as the risk items, as the two response levels, in this case, indicated significantly different levels of severity.

- Goodness of fit following threshold re-ordering: overall and item fit statistics were examined to assess whether the whole instrument and individual items fit into the Rasch model. Items with fit residuals beyond ± 2.5 and/or significant χ^2 statistics (at the 0.01 level after Bonferroni adjustment) were excluded.
- Differential item functioning (DIF): items demonstrating significant DIF (that is, responses depended on patients' age, gender, or ethnicity) were excluded from further consideration for two reasons: first because DIF can be a source of misfit in the Rasch model; second because items forming a PBM need ideally to constitute a universal measure, expressing the same aspects of HRQoL across the whole patient population (and capturing the same preferences by the valuing population), and not to distinguish significantly among sub-groups with different baseline characteristics. For this reason, although uniform DIF can be dealt with by splitting for DIF and separating the item into the different sub-groups where DIF has been identified, this was not attempted in this analysis.

Conventional psychometric tests

- Responsiveness to treatment, measured by the standardised response mean (SRM), which is estimated as the mean change in score of an item before and after treatment, divided by the standard deviation of the change score.
- Percentage of missing data
- Correlation with total CORE-OM score, expressed by Spearman's non-parametric ρ values

Results of conventional psychometric tests were used to compare the performance of items for inclusion in the final measure; thus, no formal threshold values were set to determine item exclusion.

In addition, at early stages of the analysis, expert opinion was sought to judge the appropriateness and relevance of certain CORE-OM items for inclusion in a PBM expressing peoples' perceptions on their own HRQoL relating to common mental health problems.

Items not fitting into the Rasch model (i.e. meeting one of the criteria for item exclusion described above) were excluded one at a time followed by Rasch analysis on the remaining items and subsequent testing of Rasch statistics. The person-separation index was constantly checked to ensure that the model had good ability to discriminate

amongst different respondent groups. This process was repeated until all remaining items fit into the Rasch model.

Step 2 Selecting items for the new measure

After several items were excluded, a scale fitting the Rasch model was constructed that, nevertheless, still contained a high number of items. However, there is evidence that respondents can receive, process, and remember about 7 ± 2 pieces of information, depending on the complexity of the statements [30]. For this reason, and because our purpose was to develop a concise PBM manageable in a valuation survey, further exclusion of items was undertaken, after testing different item combinations and applying the following criteria:

- The final instrument should consist of items representing the various conceptual domains of the CORE-OM. In order to keep the number of items concise, one item at maximum per domain should be selected for the final instrument.
- Overall model statistics should demonstrate best possible fit of the measure in the Rasch model, indicating the unidimensionality of the new scale.
- Response levels should be the same for all items
- Respective threshold locations for all items (the points where the probabilities of adjacent levels of response are equally likely) should ideally increase with increasing 'difficulty' of the item (as expressed by its average location). This was checked by visual inspection of the item threshold map and ensured a 'smooth' transition of responses from milder to more severe health states.
- The final instrument should be well-targeted to the patient population, covering the whole range of patients' symptom severity.

At the end of this step, an extra post hoc test was undertaken to confirm the unidimensionality of the new scale, as proposed by Smith [31]; this test has been recommended for this purpose in the Rasch literature [29, 32]. The first stage of this test is to undertake principal component analysis of the item fit residuals, in order to identify the first residual factor that primarily contributes to the variance of the data after the 'Rasch factor' has been taken into account. Subsequently, the correlation between the items and the first residual factor is examined in order to define 2 subsets of items (i.e. positively and negatively correlated). These two 'divergent' sets of items, which are most likely to breach the assumption of unidimensionality, are used to estimate two separate scores for each respondent, respectively. If the content of the whole scale is unidimensional, then each respondent should produce similar scores in the two subsets. Thus, independent *t*-tests are undertaken for each pair of scores on each respondent in order to estimate the proportion of significant tests at the

$P = 0.05$ level in the study sample. The percentage of significant tests should not exceed 5%. A confidence interval for a binomial test of proportions is calculated for the observed number of significant tests, and this value should overlap the 5% expected value for the scale to be considered unidimensional [29].

Step 3 Deriving health states for utility measurement

The item threshold map was visually inspected after all the above criteria had been satisfied, to identify the most likely item response combinations expected across the continuum of patients' symptom severity. The most likely item response combinations at each location across the scale represented frequently observed, plausible health states experienced by the study population.

Step 4 Validation of the new measure

The new measure was validated on the random sample [N400b]: the scale was tested for overall and item fit statistics and DIF. The post hoc unidimensionality test was repeated and the item threshold map was inspected to identify the most likely item response combinations in the validation sample [N400b].

Conventional psychometric tests were undertaken using SPSS 11.5 [33]. Rasch analysis was performed on RUMM2020 [34].

Results

Rasch analysis, conventional psychometric tests, and expert opinion for exclusion of items

Rasch analysis of CORE-OM on [N400a] revealed that 26 out of the 34 CORE-OM items had disordered thresholds. Threshold ordering was achieved by merging adjacent response levels following visual inspection of item category probability curves. Since no common re-scoring to all 34 items was possible to apply, we selected to use the partial-credit Rasch model for our analysis. After all thresholds were ordered, goodness of fit was assessed by examining overall model and individual item statistics. The CORE-OM did not fit into the Rasch model, with 11 items showing misfit (either a fit residual beyond ± 2.5 or a χ^2 probability significant at the 0.01 level). Moreover, 5 items demonstrated DIF.

Results of initial Rasch analysis on all items are shown in Table 2. Table 3 provides the results of conventional psychometric tests. Based on the results of Rasch analysis, a number of items were consecutively excluded from further analysis according to our exclusion criteria, until a good model fit was achieved. Conventional psychometric

test results were consulted as an extra indication of items' psychometric properties.

At an early stage of this process, it was decided to exclude items 6 (I have been physically violent to others) and 22 (I have threatened or intimidated another person). These items were judged not to be relevant to a preference-based measure, as they expressed external behaviour and not people's perceptions on their HRQoL. Moreover, both items had very low correlation with the total CORE-OM score and demonstrated low responsiveness to treatment. Item 34 was also excluded as its wording was judged to be ambiguous.

Successive Rasch analyses led to the exclusion of items 3, 8, 9, 19, 23, 24, 27, and 31 that persistently (in the initial and all consecutive analyses) misfit into the Rasch model, after considering also the results of conventional psychometric tests. For example, item 19 had relatively low responsiveness, low correlation with the initial instrument, and the highest percentage of missing data. Items 3, 8, and 31 had low correlation with CORE-OM. Item 9 had relatively low responsiveness. Items 14 and 29 were excluded because they demonstrated persistently significant DIF. Items 5, 18, 28, and 30, although did fit into the Rasch model in the initial analysis, demonstrated high fit residuals (beyond ± 2.5) at later stages and were eventually excluded from further consideration. On the other hand, items 1 and 17, which showed misfit to the model at initial stages of analysis, appeared to fit into the model at later stages, following deletion of other items, and were thus retained in the analysis.

Item 8 (I have been troubled by aches, pains, physical problems) was excluded in early stages of analysis due to misfit, which was expected, since the item expressed physical symptoms and therefore clearly belonged to a different dimension from items measuring, in their majority, emotional symptoms. Nevertheless, physical symptoms were judged to constitute an important dimension in its own right that should be captured by the final PBM; hence, although item 8 was excluded from Rasch analysis, it was decided to be combined with the final (unidimensional) product of Rasch analysis, thus creating a 2-dimensional measure tapping emotional and physical symptoms.

The 17 items of CORE-OM that fit into the Rasch model and the respective Rasch statistics are presented in Table 4. The 17-item scale had a good fit (total χ^2 probability 0.275) with an excellent ability to discriminate amongst different groups of respondents (person-separation index 0.898).

Selecting items for the emotional component of CORE-6D

The purpose of this stage of analysis was to further remove items so as to derive a concise unidimensional measure that would be manageable at a valuation exercise but at the

Table 2 Results of initial Rasch analysis of CORE-OM (all items included)

Item	Threshold	Statistics after threshold re-ordering			
		Residual	χ^2	P-value	DIF
1. I have felt terribly alone and isolated	Disordered	2.897	14.844	0.011	No
2. I have felt tense, anxious or nervous	Ordered	0.634	4.162	0.526	No
3. I have felt I have someone to turn to for support when needed	Disordered	3.273	26.580	0.000	No
4. I have felt ok about myself	Disordered	−1.339	8.616	0.125	No
5. I have felt totally lacking in energy and enthusiasm	Ordered	1.901	8.214	0.145	No
6. I have been physically violent to others	Disordered	−0.383	10.760	0.056	Yes
7. I have felt able to cope when things go wrong	Disordered	0.477	6.074	0.299	No
8. I have been troubled by aches, pains, physical problems	Disordered	3.614	23.730	0.000	Yes
9. I have thought of hurting myself	Disordered	−1.470	22.802	0.000	No
10. Talking to people has felt too much for me	Disordered	−0.138	1.198	0.945	No
11. Tension/anxiety have prevented me doing important things	Disordered	0.081	1.095	0.955	No
12. I have been happy with the things I have done	Disordered	0.306	0.319	0.997	No
13. I have been disturbed by unwanted thoughts and feelings	Disordered	1.607	16.574	0.005	No
14. I have felt like crying	Ordered	−0.876	2.116	0.833	Yes
15. I have felt panic or terror	Disordered	−0.121	3.630	0.604	No
16. I made plans to end my life	Disordered	−1.141	13.536	0.019	No
17. I have felt overwhelmed by my problems	Ordered	−2.645	13.646	0.018	No
18. I have had difficulty of getting to sleep or staying asleep	Disordered	0.688	8.987	0.110	No
19. I have felt warmth or affection for someone	Disordered	4.806	54.235	0.000	No
20. My problems have been impossible to put to one side	Ordered	0.299	0.473	0.993	No
21. I have been able to do most things I needed to	Disordered	0.904	10.794	0.056	No
22. I have threatened or intimidated another person	Disordered	1.357	9.625	0.087	No
23. I have felt despairing or hopeless	Disordered	−4.333	37.877	0.000	No
24. I have thought it would be better if I were dead	Disordered	−1.867	20.908	0.001	Yes
25. I have felt criticised by other people	Disordered	−0.262	3.023	0.696	No
26. I have thought I have no friends	Disordered	−0.073	5.034	0.412	No
27. I have felt unhappy	Ordered	−4.101	25.772	0.000	No
28. Unwanted images or memories have been distressing me	Disordered	−0.440	12.087	0.034	No
29. I have been irritable when with other people	Ordered	2.706	13.455	0.019	Yes
30. I have thought I am to blame for my problems & difficulties	Ordered	0.847	3.083	0.687	No
31. I have felt optimistic about my future	Disordered	3.779	20.251	0.001	No
32. I have achieved the things I wanted to	Disordered	0.025	6.792	0.237	No
33. I have felt humiliated or shamed by other people	Disordered	1.504	11.613	0.040	No
34. I have hurt myself physically or taken risks with my health	Disordered	0.675	3.719	0.591	No

Overall model statistics after threshold re-orderingTotal item $\chi^2 = 425.624$; $P = 0.000$

Person-separation index: 0.93

Residuals >2.5 or < -2.5 are considered high; $P < 0.01$ indicates items that do not meet Rasch item fit criteria. All statistics showing item misfit into the Rasch model are illustrated in bold

same time would capture major conceptual domains of CORE-OM.

All items belonging to the conceptual domains ‘symptoms—depression’, ‘symptoms—physical’, and ‘risk/harm to others’ had already been excluded at a previous stage.

Expert judgement concluded that the conceptual domains ‘symptoms—anxiety’ (represented by items 2, 11, 15, 20 in the 17-item scale), ‘functioning—general’ (items 7, 12, 21, 32 in the 17-item scale), ‘functioning—close relationships’ (items 1, 26 in 17-item scale), ‘functioning—social

Table 3 Results of conventional psychometric tests on CORE-OM

Item	SRM	Missing data (%)	Spearman's ρ value
1. I have felt terribly alone and isolated	0.99	0.4	0.714
2. I have felt tense, anxious, or nervous	1.18	0.3	0.603
3. I have felt I have someone to turn to for support when needed	0.65	0.7	0.419
4. I have felt ok about myself	1.00	0.6	0.646
5. I have felt totally lacking in energy and enthusiasm	0.96	0.4	0.587
6. I have been physically violent to others	0.24	0.5	0.282
7. I have felt able to cope when things go wrong	0.78	0.6	0.594
8. I have been troubled by aches, pains, physical problems	0.61	0.7	0.276
9. I have thought of hurting myself	0.46	0.4	0.531
10. Talking to people has felt too much for me	0.81	0.7	0.548
11. Tension/anxiety have prevented me doing important things	0.89	0.8	0.642
12. I have been happy with the things I have done	0.85	0.8	0.624
13. I have been disturbed by unwanted thoughts and feelings	0.95	0.5	0.564
14. I have felt like crying	1.19	0.3	0.630
15. I have felt panic or terror	0.84	0.4	0.576
16. I made plans to end my life	0.29	1.0	0.436
17. I have felt overwhelmed by my problems	1.09	1.0	0.744
18. I have had difficulty of getting to sleep or staying asleep	0.93	0.6	0.521
19. I have felt warmth or affection for someone	0.33	2.4	0.299
20. My problems have been impossible to put to one side	1.04	0.9	0.629
21. I have been able to do most things I needed to	0.69	0.8	0.568
22. I have threatened or intimidated another person	0.32	1.0	0.272
23. I have felt despairing or hopeless	1.09	0.8	0.785
24. I have thought it would be better if I were dead	0.58	0.7	0.647
25. I have felt criticised by other people	0.70	0.8	0.558
26. I have thought I have no friends	0.65	0.9	0.595
27. I have felt unhappy	1.26	0.5	0.731
28. Unwanted images or memories have been distressing me	0.89	0.6	0.576
29. I have been irritable when with other people	0.86	0.9	0.554
30. I have thought I am to blame for my problems & difficulties	0.80	0.5	0.533
31. I have felt optimistic about my future	0.81	1.0	0.465
32. I have achieved the things I wanted to	0.86	1.5	0.590
33. I have felt humiliated or shamed by other people	0.61	1.1	0.557
34. I have hurt myself physically or taken risks with my health	0.27	0.9	0.348

SRM standardised response mean; Spearman's ρ value expresses correlation with total CORE-OM score. In bold: SRM values <0.50; % of missing data \geq 1.0%; and Spearman's ρ values <0.500

relationships' (items 10, 25, 33 in the 17-item scale) and 'risk/harm to self' (item 16 in the 17-item scale) reflected major domains in people with common mental health problems and should be represented in the final construct. The 'subjective well-being' domain (items 4 and 17 in the 17-item scale) covered the overall perception of person's HRQoL rather than distinct symptoms/problems of the study population. Indeed, this domain had been previously found to highly correlate with items in the overall 'problems' domain [20]. Regarding the 'symptoms—trauma' domain (item 13 in the 17-item scale), this was considered less relevant for this HRQoL measure. More importantly,

attempts to include items of these last two domains ('subjective well-being' and 'symptoms—trauma') in the final measure resulted in a scale not satisfying the criterion of 'smooth' transition of the response thresholds from milder to more severe health states, set in step 2. Consequently, it was decided to exclude these domains from the final measure.

Items were excluded one at a time and Rasch statistics as well as the person separation index were constantly checked. Finally, various combinations of 5 items (of those included in the 17-item scale), one from each of the 5 CORE-OM conceptual domains considered for the

Table 4 Results of Rasch analysis with the 17 items of CORE-OM fitting into the Rasch model

Item	Threshold	Item statistics			
		Residual	χ^2	<i>P</i> -value	DIF
1. I have felt terribly alone and isolated	Ordered	1.415	10.118	0.072	No
2. I have felt tense, anxious or nervous	Ordered	−0.373	2.658	0.752	No
4. I have felt ok about myself	Ordered	−0.107	2.326	0.802	No
7. I have felt able to cope when things go wrong	Ordered	0.371	5.829	0.323	No
10. Talking to people has felt too much for me	Ordered	0.546	4.614	0.465	No
11. Tension/anxiety have prevented me doing important things	Ordered	−0.191	6.021	0.304	No
12. I have been happy with the things I have done	Ordered	0.708	1.848	0.870	No
13. I have been disturbed by unwanted thoughts and feelings	Ordered	2.376	10.195	0.070	No
15. I have felt panic or terror	Ordered	0.133	5.590	0.348	No
16. I made plans to end my life	Ordered	−0.485	4.897	0.428	No
17. I have felt overwhelmed by my problems	Ordered	−2.084	11.369	0.045	No
20. My problems have been impossible to put to one side	Ordered	0.254	1.877	0.866	No
21. I have been able to do most things I needed to	Ordered	1.424	3.410	0.637	No
25. I have felt criticised by other people	Ordered	0.918	3.362	0.644	No
26. I have thought I have no friends	Ordered	0.742	8.993	0.110	No
32. I have achieved the things I wanted to	Ordered	0.799	1.426	0.921	No
33. I have felt humiliated or shamed by other people	Ordered	−0.899	7.809	0.167	No
Overall model statistics		Total item $\chi^2 = 92.342$; <i>P</i> = 0.275			
		Person-separation index: 0.898			

Residuals >2.5 or < −2.5 are considered high; *P* < 0.01 indicates items that do not meet Rasch item fit criteria. All items were shown to fit into the Rasch model

emotional component of the new measure, were tested for their fit into the Rasch model, in order to construct a final scale that would meet the set criteria for this step.

Testing of various item combinations resulted in a measure consisting of 5 items (1, 15, 16, 21, 33), each with 3 levels of response, common to all items (‘not at all’, ‘only occasionally or sometimes’, and ‘often, most or all the time’). The 5 items belonged to 5 major CORE-OM conceptual domains, respectively. The scale demonstrated good model fit (χ^2 probability 0.69). All items fit into the model, as shown in Table 5; no DIF was observed. The person-separation index reached 0.659. Respective threshold

locations increased with increasing item difficulty, as shown in the item threshold map (Fig. 2). The item map demonstrates that the instrument is well targeted to the study population as it is able to capture the whole range of severity of mental symptoms, with minimal floor or ceiling effects and good spread of items across the full range of respondents’ scores (Fig. 3).

According to the results of the post hoc test proposed by Smith [31], the proportion of independent *t*-tests that were significant at the 0.05 level was 1.34% (well below 5%), thus confirming the unidimensionality of the emotional component of CORE-6D.

Table 5 Rasch statistics of the emotional component of CORE-6D

Item	Conceptual domain	Rasch analysis statistics		
		Residual	χ^2	<i>P</i> -value
1. I have felt terribly alone and isolated	Functioning—close relationships	−0.099	2.044	0.843
15. I have felt panic or terror	Symptoms—anxiety	−0.058	3.403	0.638
16. I made plans to end my life	Risk/harm to self	−0.358	5.812	0.325
21. I have been able to do most things I needed to	Functioning—general	0.717	6.520	0.259
33. I have felt humiliated or shamed by other people	Functioning—social relationships	0.156	3.191	0.671
Overall model statistics		Total item $\chi^2 = 20.970$; <i>P</i> = 0.694		
		Person-separation index: 0.659		

Fig. 2 Item threshold map of the emotional component of the CORE-6D illustrating the plausible health states obtained by Rasch analysis

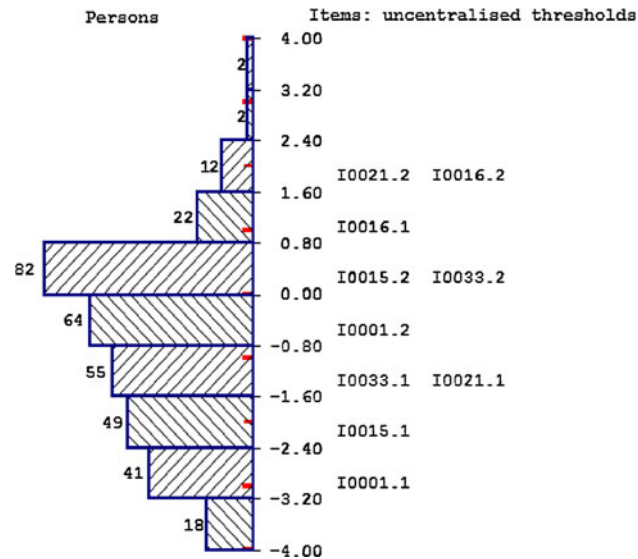
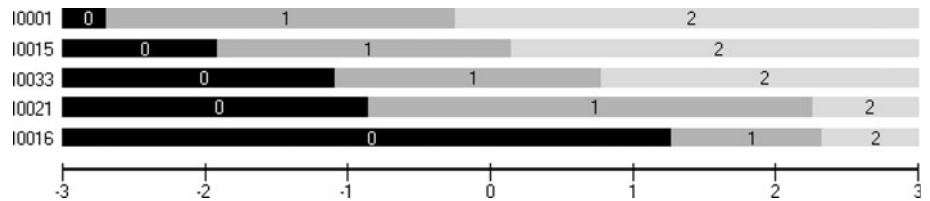


Fig. 3 Item map of the emotional component of the CORE-6D

Deriving plausible health states from the emotional component of CORE-6D for utility measurement

Derivation of plausible health states was based on the item threshold map (Fig. 2). The map illustrates the most likely combinations of item responses expected to be obtained by the study population at various levels (locations) of symptom severity. Items have been ordered from the easiest (item 1) to the most difficult one (item 16), as indicated by their average location in the Rasch model. Shaded areas 0 (black), 1 (dark grey) and 2 (light grey) correspond to the

3 response levels ‘not at all’, ‘only occasionally or sometimes’, and ‘often, most or all the time’, respectively, with the exception of item 21, which is positively worded and therefore response levels are reversed. Threshold locations between response levels 0–1 and 1–2 increase (that is, they move from the left to the right) with increasing difficulty of the item, thus ensuring a smooth transition of responses from milder to more severe symptoms. The item threshold map allows prediction of the most likely responses at various levels of severity. For example, a person whose symptom severity corresponds to location +1 on the logit scale is expected to most likely respond 22210 (to items 1, 15, 33, 21 and 16, respectively).

Each combination of item responses represents a plausible health state, likely to be observed in people with common mental health problems. As illustrated in Table 6, 11 distinct health states can be identified. These states covered 37% of complete responses in N400a. In contrast, the coverage of health states derived using an orthogonal block design on the full range of $3^5 = 243$ potential health states of the emotional component of CORE-6D was only 7%. Moreover, some of the states generated using the latter approach were not credible, as, for example, they described a situation where a person ‘never felt alone and isolated’ and at the same time ‘made plans to end their life often, most or all the time’.

Validation of the emotional component of CORE-6D

The emotional component of CORE-6D was validated on the random sample [N400b]: the scale had satisfactory

Table 6 Health states of the emotional component of CORE-6D as identified by the item threshold map

Item	Health states										
	1	2	3	4	5	6	7	8	9	10	11
1. I have felt terribly alone and isolated	N	S	S	S	S	O	O	O	O	O	O
15. I have felt panic or terror	N	N	S	S	S	S	O	O	O	O	O
33. I have felt humiliated or shamed by other people	N	N	N	S	S	S	S	O	O	O	O
21. I have been able to do most things I needed to	O	O	O	O	S	S	S	S	S	N	N
16. I made plans to end my life	N	N	N	N	N	N	N	N	S	S	O

N not at all; S only occasionally or sometimes; O often, most or all the time; note that item 21 is positively worded and therefore response levels are reversed

overall and item fit statistics and no DIF was observed. The post hoc unidimensionality test verified the scale's unidimensionality in this sample, too, and the item threshold map indicated the same most likely item response combinations (reflecting plausible health states) with those demonstrated by the analyses on sample [N400a].

Discussion and conclusions

This paper proposes a methodology that uses mainly Rasch analysis to develop plausible health states from existing CSMS that have no clear multidimensional structure; in such cases, conventional approaches for generating states from health state classifications (e.g. orthogonal block designs) are not appropriate, as, by treating items as independent (uncorrelated) statements, they are likely to result in formation of implausible health states. In contrast, the proposed 'Rasch vignette approach' helps create credible health states comprising combinations of item responses observed in a real population. Indeed, the health states developed with this method represent not only plausible, but also the most likely combinations of responses over a continuum of symptom severity, thus allowing prediction of a person's severity of symptoms based on his/her responses and vice versa. On the other hand, in their clustering-based approach, Sugar et al. [17] combined the most frequent individual item responses within each cluster in order to develop health state descriptions. However, the resulting item response combinations were not necessarily the most frequently observed in the study sample; what's more, they might have not been observed at all in the sample.

One limitation of our approach, similar to the methodology proposed by Sugar et al. [17], is that the number of generated health states is limited and does not capture the whole range of plausible combinations of responses. In the case of the emotional component of CORE-6D, the Rasch vignette approach generated 11 health states, which, nevertheless, covered 37% of the study sample's complete responses; on the other hand, use of an orthogonal block design, which assumes that items are independent statements, achieved a much lower coverage of 7%, and, more importantly, generated a number of implausible health states.

Despite generating a limited number of health states, application of our approach allows the valuation of all potential health states described by CORE-6D: an advantage of Rasch analysis over the clustering-based approach is that it assigns all potential health states (i.e. all combinations of item responses including those not illustrated in item threshold maps) to different locations along the scale

according to their level of severity. The relationship between the health states' location across the latent variable and the respective utility values obtained in a valuation exercise can be estimated and used to generate utility values for all patients completing CORE-OM. This solution has been explored, using regression techniques, in a subsequent application of this approach on the Flushing questionnaire [35]. The findings of this latter study show that it is possible to assign appropriate utility values to all potential health states of a measure based on their location along the latent variable as estimated by Rasch analysis.

The emotional component of CORE-6D comprises a unidimensional 5-item scale, able to capture the full range of severity of emotional symptoms in people with common mental health problems. The person-separation index of this scale was approximately 0.66, which is somewhat lower than the 0.70 value that is generally considered acceptable for group comparison [36]. Nevertheless, the 0.66 figure was deemed adequate for our purpose, which was the development of a PBM, considering that the ability of the scale to discriminate amongst different respondent groups needed to be traded off with its conciseness and convenience in a valuation survey, where respondents need to process a combination of individual statements rather than a summated scale score.

The proposed Rasch vignette approach has led to identification of 11 plausible health states. These states, combined with 3 response levels (same as those of the 5 'emotional' items) of item 8 of the original CORE-OM (I have been troubled by aches, pains, or physical problems), produce a 2-dimensional set of $11 \times 3 = 33$ plausible health states that can be used to value the overall emotional and physical HRQoL in people with common mental health problems. The next step of this study, recently completed, was to undertake a valuation survey in a representative sample of the UK population, in order to attach appropriate utility values to all health states of CORE-6D and thus convert it into a preference-based index. This new condition-specific PBM is suitable to use in the area of mental health, where the use of generic PBMs such as EQ-5D has been shown to be problematic [27, 37, 38]. Since this measure has been derived from CORE-OM, an instrument routinely used for outcome monitoring in people with common mental health problems in the United Kingdom, it is expected that this study will enable wider assessment of healthcare interventions for the management of common mental health problems in the form of cost-utility analysis.

Acknowledgments Ifigeneia Mavranzouli undertook this work as a PhD student in the School of Health and Related Research at the University of Sheffield. We are grateful to the UK MRC Methodology Programme for supporting this work and to the CORE System Trustees for giving approval for the development of the CORE-6D.

References

1. Brazier, J., Ratcliffe, J., Salomon, J. A., & Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford/New York: Oxford University Press.
2. Dolan, P. (1997). Modelling valuations for EuroQol health states. *Medical Care*, 35(11), 1095–1108.
3. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference based measure of health from the SF-36. *Journal of Health Economics*, 21(2), 271–292.
4. Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., et al. (2002). Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care*, 40(2), 113–128.
5. Brazier, J., Deverill, M., Green, C., Harper, R., & Booth, A. (1999). A review of the use of health status measures in economic evaluation. *Health Technology Assessment*, 3(9), 1–164.
6. Tsuchiya, A., Brazier, J., McColl, E., & Parkin, D. (2002). Deriving preference-based single indices from non-preference based condition-specific instruments: Converting AQLQ into EQ5D indices. Sheffield Health Economics Group Discussion Paper 02/01. Sheffield: School of Health and Related Research, University of Sheffield. Available from <http://www.shef.ac.uk/scharr/sections/heds/discussion.html>.
7. McKenzie, L., & Van der Pol, M. (2009). Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: The potential to estimate QALYs without generic preference data. *Value in Health*, 12(1), 167–171.
8. Gray, A. M., Rivero-Arias, O., & Clarke, P. M. (2006). Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making*, 26(1), 18–29.
9. Brazier, J. E., Yang, Y., Tsuchiya, A., & Rowen, D. L. (2010). A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *European Journal of Health Economics*, 11(2), 215–225.
10. Brazier, J. E., Roberts, J., Platts, M., & Zoellner, Y. F. (2005). Estimating a preference-based index for a menopause specific health related quality of life questionnaire. *Health and Quality of Life Outcomes*, 3, 13.
11. Yang, Y., Tsuchiya, A., Brazier, J. E., & Young, T. A. (2007). Estimating a preference-based single index from the asthma quality of life questionnaire (AQLQ). Health Economics and Decision Science Discussion paper No. 07/02. Sheffield: School of Health and Related Research. Available from <http://www.shef.ac.uk/scharr/sections/heds/discussion.html>.
12. Brazier, J., Czoski-Murray, C., Roberts, J., Brown, M., Symonds, T., & Kelleher, C. (2008). Estimation of a preference-based index from a condition-specific measure: The king's health questionnaire. *Medical Decision Making*, 28(1), 113–126.
13. Yang, Y., Brazier, J., Tsuchiya, A., & Coyne, K. (2009). Estimating a preference-based single index from the overactive bladder questionnaire. *Value in Health*, 12(1), 159–166.
14. Chatfield, C., & Collins, A. J. (1980). *Introduction to multivariate analysis*. Cambridge: Chapman and Hall, University Press.
15. Young, T. A., Yang, Y., Brazier, J. E., & Tsuchiya, A. (2010). The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: The case of moving from AQLQ to AQL-5D. *Medical Decision Making*. doi:10.1177/0272989X10364846.
16. Young, T., Yang, Y., Brazier, J. E., Tsuchiya, A., & Coyne, K. (2009). The first stage of developing preference-based measures: Constructing a health-state classification using Rasch analysis. *Quality of Life Research*, 18(2), 253–265.
17. Sugar, C. A., Sturm, R., Lee, T. T., Sherbourne, C. D., Olshen, R. A., Wells, K. B., et al. (1998). Empirically defined health states for depression from the SF-12. *Health Services Research*, 33(4 Pt 1), 911–928.
18. Bond, T. G., & Fox, C. M. (2006). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.
19. Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., et al. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. Clinical outcomes in routine evaluation-outcome measures. *Journal of Consulting and Clinical Psychology*, 69(2), 184–196.
20. Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., et al. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, 180, 51–60.
21. Barkham, M., Gilbert, N., Connell, J., Marshall, C., & Twigg, E. (2005). Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. *British Journal of Psychiatry*, 186, 239–246.
22. Evans, C., Connell, J., Barkham, M., Marshall, C., & Mellor-Clark, J. (2003). Practice-based evidence: Benchmarking NHS primary care counselling services at national and local levels. *Clinical Psychology & Psychotherapy*, 10, 374–388.
23. Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M., et al. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology*, 47(Pt 4), 397–415.
24. Jacobs, R. (2009). Investigating patient outcome measures in mental health. CHE Research Paper Number 48. York: Centre for Health Economics, University of York.
25. Chisholm, D., Healey, A., & Knapp, M. (1997). QALYs and mental health care. *Social Psychiatry and Psychiatric Epidemiology*, 32(2), 68–75.
26. Knapp, M., & Mangalore, R. (2007). “The trouble with QALYs...”. *Epidemiologia e Psichiatria Sociale*, 16(4), 289–293.
27. Brazier, J. (2008). Measuring and valuing mental health for use in economic evaluation. *Journal of Health Services Research & Policy*, 13(Suppl 3), 70–75.
28. Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8, 33.
29. Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism (Arthritis Care & Research)*, 57(8), 1358–1362.
30. Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
31. Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205–231.
32. Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, 20(1), 1048–1051. Available from <http://www.rasch.org/rmt/rmt201c.htm>.
33. SPSS for Windows ©, 11.5 (2002). Chicago: SPSS Inc.
34. Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). RUMM2020. Perth: RUMM Laboratory Pty Ltd. <http://www.rummlab.com>.

35. Young, T. A., Rowen, D., Norquist, J., & Brazier, J. E. (2010). Developing preference-based health measures: Using Rasch analysis to generate health state values. *Quality of Life Research*, *19*(6), 907–917.
36. Fischer, W., Jr. (1992). Reliability statistics. *Rasch Measurement Transactions*, *6*(3), 238. Available from <http://www.rasch.org/rmt/rmt63i.htm>.
37. Moock, J., & Kohlmann, T. (2008). Comparing preference-based quality-of-life measures: Results from rehabilitation patients with musculoskeletal, cardiovascular, or psychosomatic disorders. *Quality of Life Research*, *17*(3), 485–495.
38. van de Willige, G., Wiersma, D., Nienhuis, F. J., & Jenner, J. A. (2005). Changes in quality of life in chronic psychiatric patients: A comparison between EuroQol (EQ-5D) and WHOQoL. *Quality of Life Research*, *14*(2), 441–451.