Health Economics and Decision Science, School of Health and Related Research (ScHARR)

The derivation of a preference-based measure for people with common mental health problems from the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM)

Ifigeneia Mavranezouli

Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

January 2014

Author's declaration

I declare that this thesis is my original work and that the contents and views expressed are my own. I have conducted all the work reported in this thesis, with support from those mentioned under "Acknowledgements". The selection of the 22 health states of CORE-6D that were included in the valuation survey reported in Chapter 6 was made in collaboration with Donna Rowen. Figure 17 in the same chapter has been produced by Donna Rowen.

Funding

Funding for the valuation survey undertaken as part of this thesis was provided by the MRC-NIHR Methodology Research Programme (project number 06/97/04).

Acknowledgements

I would like to thank my supervisors John Brazier and Michael Barkham for their ongoing support, advice and direction throughout the 8 years I have been conducting this work.

Special thanks to Donna Rowen for advice on statistical issues and useful comments on various stages of this work, especially in the valuation of CORE-6D and the assessment of its performance. Thanks also to Tracey Young for providing expert advice in the development of the CORE-6D health state classification, to Rachel Ibbotson for overviewing and managing the primary data collection for the valuation survey, and to Iftekhar Khan and Stephen Morris for advice in analysing the valuation data. I would also like to thank all the interviewees who took part in the valuation survey.

I want to say a big thank you to Stephen Pilling for endless support and continuous encouragement to start, carry on, and complete this thesis, especially through stressful times. Your support and trust in my abilities has been inspiring! Many thanks also to Tim Kendall and all my colleagues at the National Collaborating Centre for Mental Health for their patience, understanding and support whenever things became too hectic or stressful for me. In particular, I would like to thank Sarah Stockton for her time and advice on systematic searching and Clare Taylor for advice and several tips on editing the thesis.

Also, thanks to Janice Connell for preparing and providing the CORE-OM datasets that were analysed in order to develop and validate the CORE-6D health state classification; Jane Cahill and Jane Morrell for providing the datasets that were used for the assessment of the psychometric properties of CORE-6D, as well as for valuable information and advice on analysis of the data.

Finally, a million thanks to my family for putting up with my stress and my heavy workloads. My partner Ilias for infinite patience, invaluable support and continuous encouragement especially when things felt unmanageable. My children Iasonas and Nafsika for being patient and hardly ever complaining when mummy was busy doing her ‘homework’. Apologies for not being around

as much as you would have wanted over the past years... My parents for always being there and offering to help in any way they could, in particular looking after the children at stressful and busy periods. And my brother, Simos, for being a computer geek and responding instantly to any last-minute emergency request! Without support from my family, completion of this work would never have been possible.

Abstract

Background: Generic preference-based measures (PBMs), such as the EQ-5D and SF-6D, are widely used for the estimation of Quality Adjusted Life Years in cost-utility analyses of healthcare interventions. However, their relevance in some disease areas, including mental health, has been questioned.

Objective of the thesis: To derive a PBM specific to mental health problems from an existing condition-specific measure (CSM)

Methods: A systematic literature review was conducted to identify an appropriate CSM for the derivation of a health state classification. Derivation of the new measure was achieved using novel methodology developed for this purpose, due to the high correlation across the items of the original CSM. Selected health states were valued by members of the public. Regression analysis was employed to predict utility values for all states of the health state classification. Psychometric and qualitative assessments evaluated the performance of the new PBM compared with generic PBMs and the original CSM.

Results: The Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) was selected as the basis for the derivation of the new PBM. Application of novel methodology based primarily on Rasch analysis resulted in the development of CORE-6D, a health state classification that consists of a 5-item emotional component and a physical item. Rasch analysis was used to select plausible health states for valuation. A highly predictive regression model was used to attach utility values to all CORE-6D health states. The new PBM has shown promising results regarding its psychometric properties compared with generic PBMs and suffers from little loss of information relative to the original measure, CORE-OM. Further research needs to validate these findings.

Conclusion: The CORE-6D preference-based index will enable cost-utility analysis of mental health interventions using existing and prospective CORE-OM datasets. The new methodology for deriving PBMs from existing instruments can be useful for the derivation of PBMs from other instruments with highly correlated dimensions.

Related publications and conference presentations

Publications in peer-reviewed journals

1. Mavranezouli I, Brazier JE, Rowen D, Barkham M. **Estimating a Preference-Based Index from the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM): valuation of CORE-6D.** Medical Decision Making 2013; 33(3): 381-395

2. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, Barkham M, Ibbotson R. **Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient-based measures of outcome).** Health Technology Assessment 2012; 16(32): 1-114

3. Mavranezouli I, Brazier JE, Young TA, Barkham M. **Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems (CORE-OM).** Quality of Life Research 2011; 20: 321-333

Related conference presentations

1. Young T, Brazier J, Rowen D, Mulhern B, Mavranezouli I. Deriving **preference-based utility measures from existing measures: How Health Economists Make Use of Rasch Analysis.**
   6th UK Rasch User Group Meeting, Leeds (UK), 20 March 2012

2. Mavranezouli I, Brazier JE, Rowen D, Barkham M. **The development of a condition-specific preference-based measure for common mental health disorders from the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM) using Rasch analysis.**
   Presented at the Health Economists' Study Group (HESG) 2010 Summer Conference, Cork (Ireland), 23-25 June 2010

Contents

List of tables

List of figures

List of abbreviations

AQLQ	Asthma Quality of Life Questionnaire
ADDQoL	Audit of Diabetes-Dependent Quality-of-Life
ANOVA	Analysis of Variance
BDI	Beck Depression Inventory
BPRS	Brief Psychiatric Rating Scale
CAMPHOR	Cambridge Pulmonary Hypertension Outcome Review
CAN	Camberwell Assessment of Needs
CANSAS	Camberwell Assessment of Needs Short Appraisal Schedule
CIS-R	Clinical Interview Schedule - Revised
CORE-6D	Clinical Outcomes in Routine Evaluation – 6-dimensional health state classification
CORE-OM	Clinical Outcomes in Routine Evaluation - Outcome Measure
CoSMeQ	Condition-Specific Methodology for estimating QALYs
CSM	Condition-Specific Measure
DCE	Discrete Choice Experiment
DIF	Differential Item Functioning
DUI	Diabetes Utility Index
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire
EPDS	Edinburgh Postnatal Depression Scale
ES	Effect Size
FA	Factor Analysis
FACE	Functional assessment of the Care Environment
GAD	Generalised Anxiety Disorder
GAF	Global Assessment of Functioning
GAS	Global Assessment Scale
GHQ	General Health Questionnaire
GP	General Practitioner
HAQ	Health Assessment Questionnaire
HAM-D	Hamilton Depression Rating Scale

HEDS	Health Economics and Decision Science
HoNOS	Health of the Nation Outcome Scales
HRQoL	Health-Related Quality of Life
HUI	Health Utilities Index
IAPT	Improving Access to Psychological Therapies
ICC	Intraclass Correlation Coefficient
ICER	Incremental Cost Effectiveness Ratio
CORE IMS	CORE Information Management Systems
IPSS	International Prostate Symptom Score
IRT	Item Response Theory
KHQ	King's Health Questionnaire
MAE	Mean Absolute Error
MAUT	Multi-Attribute Utility Theory
MSIS-29	Multiple Sclerosis Impact Scale 29
MVH	Measurement and Valuation of Health
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
OAB-q	Overactive Bladder Questionnaire
OCD	Obsessive Compulsive Disorder
OLS	Ordinary Least Squares
PANSS	Positive and Negative Syndrome Scale
PBM	Preference-Based Measure
PCA	Principal Components Analysis
PHASE	Psychological Health by Assessing Self-help Education
PHQ-9	Patient Health Questionnaire - 9 items
PMS	Psychiatric Morbidity Survey
PoNDER	Postnatal Depression Effectiveness Randomised Controlled Trial
PROM	Patient-Reported Outcome Measure
PSI	Person Separation Index
PTRC	Psychological Therapies Research Centre
QALY	Quality Adjusted Life Year
QWB	Quality of Well-Being scale
RCT	Randomised Controlled Trial

RMSE	Root Mean Squared Error
ScHARR	School of Health And Related Research
SCL-90-R	Symptom Checklist-90-Revised
SF-36	Short Form (36) Health Survey
SG	Standard Gamble
SQOL	Sexual Quality of Life questionnaire
SRM	Standardised Response Mean
TAG	Threshold Assessment Grid
TTO	Time Trade-Off
VAS	Visual Analogue Scale
WHO	World Health Organization

Overview [executive summary]

Background: Economic evaluation of healthcare interventions in the form of cost-utility analysis is increasingly advocated by regulatory bodies worldwide, including the National Institute for Health and Care Excellence (NICE) in England and Wales. The most commonly used outcome measure in this type of analysis is the Quality Adjusted Life Year (QALY). Generic preference-based measures (PBMs), such as the EQ-5D, the SF-6D and the HUI-3 are widely used for the estimation of QALYs. These measures consist of a health state classification describing Health-Related Quality of Life (HRQoL), and an algorithm converting the HRQoL in each health state into a utility value, based on public preferences elicited in a valuation survey. Despite their wide use, generic measures of health may be inappropriate or insensitive in capturing relevant aspects of HRQoL in some medical conditions. On the other hand, condition-specific measures (CSMs) of outcome focus on specific symptoms and aspects of HRQoL characterising a disease area, and therefore are expected to be more relevant and sensitive than generic measures in capturing the impact of the disease on patients' HRQoL. However, the majority of the available CSMs are not preference-based, and therefore are not suitable for calculation of QALYs. Over the last years, there has been an increased interest in the development of PBMs directly from existing CSMs. One area where concerns about the relevance and sensitivity of generic measures have been expressed is mental health, leading to proposals for the development of a mental health-specific 'generic' PBM that can be used across the full spectrum of mental disorders.

Aims and objectives: The aim of this thesis was the derivation of a PBM that is relevant to people with mental health problems from an existing CSM that is currently used in mental health research and practice. Specific objectives of the thesis were as follows:

1. To assess the use and psychometric performance of generic PBMs in mental health research and practice, in order to explore in depth the appropriateness of using such measures in cost-utility analyses

conducted in the area of mental health and to confirm the need for a new PBM developed specifically for this area

2. To identify an appropriate CSM used in the area of mental health with proven validity, sensitivity and acceptability that is able to capture a wide range of symptoms and HRQoL aspects that are relevant to people with mental disorders, to use as the basis for the derivation of a PBM specifically designed for use in mental health, in particular within the UK National Health Service (NHS) context.
3. To derive a new health state classification from the selected CSM and subsequently attach appropriate utility values to all health states described by the new measure, so as to develop a PBM for people with mental health problems.
4. To evaluate the performance of the new PBM relative to generic ones, the loss of information relative to the original CSM which it was derived from, and its relevance to people with mental health problems.

Methods and results: In order to fulfil the specific objectives of the thesis, four pieces of work were undertaken. An overview of methods and results is provided separately for each piece of work:

1. Assessment of the use and psychometric performance of generic PBMs in mental health research and practice

Methods: A systematic literature review of the use and psychometric performance of the EQ-5D, SF-6D and HUI-3 in mental health research and practice was undertaken. The review included literature reviews that reported data on the use and psychometric properties of generic PBMs in adults with a primary diagnosis of a mental disorder.

Results: EQ-5D and SF-6D perform satisfactorily in depression and, to a lesser extent, in anxiety and personality disorders. Results were mixed in schizophrenia and bipolar disorder. Results suggest that EQ-5D may be picking depressive symptoms (or comorbid depression) rather than core symptoms associated with a range of conditions, including anxiety and schizophrenia. No reviews on the use and properties of HUI-3 in the area of mental health were identified. A review of qualitative evidence revealed that

generic PBMs fail to address the complexity of quality of life measurement and the broad range of domains that are important to people with mental health problems. Evidence also suggests that mental health professionals are rather reluctant to using generic measures for the measurement of HRQoL in clinical research and practice, as CSMs are deemed more appropriate for and sensitive to capturing relevant aspects of HRQoL.

2. Identification of an appropriate CSM to be used as the basis for the derivation of a PBM specific to mental health problems

Methods: A systematic literature review on outcome measures used in mental health research and practice was carried out. The review included literature reviews reporting on the use and properties of instruments used for outcome measurement and monitoring of adults with mental disorders, including symptoms, functioning and HRQoL. The most appropriate CSM of those identified in the systematic review was selected for the derivation of the new PBM, based on a number of considerations:

- Broad coverage of symptoms and aspects of HRQoL, including both mental and physical health aspects
- Psychometric properties: established construct validity and responsiveness
- Wide coverage within the British NHS
- Applicability across primary and secondary settings
- Free use
- Patient-reported
- Applicability across a range of mental disorders

Results: The review of reviews identified a wide range of instruments used for outcome measurement and monitoring of people with mental health problems. Of these, the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) and the Health of the Nation Outcome Scales (HoNOS) appeared to fulfil all or most of the set criteria that were used to determine the

appropriateness of a CSM to serve as the basis for the derivation of a new PBM that is specific to mental health.

HoNOS is a valid, reliable and responsive outcome measure. It can be used for free within the British NHS and is widely used in the UK clinical practice. HoNOS has not been designed for use in primary care. It is a measure of severe and enduring mental illness and as such it may not be appropriate for use in people with mild and moderate mental illness. Moreover, HoNOS is clinician-rated, whereas PBMs are traditionally patient-reported, and this was considered a major disadvantage against its use as the basis for the derivation of a mental health-specific PBM.

CORE-OM is a valid and responsive measure with wide coverage within the British NHS. It is free to use and is applicable across primary and secondary settings. In contrast to HoNOS, CORE-OM is patient-reported, which is a big advantage for its use in the derivation of a mental health-specific PBM. The CORE-OM has been designed for outcome measurement in people with common mental health problems such as depression and anxiety disorders, which, nevertheless, are the most prevalent mental disorders in the UK. Qualitative assessment indicated that the CORE-OM items tap the majority of areas of HRQoL that are considered important by people with mental health problems. Based on these criteria, CORE-OM was selected as the basis for the derivation of a new PBM that is relevant to people with common mental health problems.

3. Derivation of a new health state classification, valuation and modelling of utility values, leading to the development of a new PBM for people with mental health problems

Methods: A further systematic review was conducted, to explore and assess the different methods that have been reported in the literature for the derivation of health state classifications that are amenable to valuation from existing, longer measures (either generic or condition-specific), aiming to identify appropriate methodologies that might contribute to the derivation of a new

health state classification from the CORE-OM. However, due to the high correlation across the domains of CORE-OM, which precluded the use of standard methodology for the selection of dimensions and items for the health state classification, a novel methodology was developed and applied, which was primarily based on Rasch analysis, supplemented by a range of psychometric tests undertaken on each item of CORE-OM, including acceptability, degree of floor and ceiling effects, responsiveness to change over time, and correlation with the CORE-OM. Following the development of the health state classification, a valuation survey of 220 members of the public in South Yorkshire was subsequently undertaken using the time trade-off (TTO) method. The selection of the health states valued in the survey was also based on Rasch analysis. Finally, regression analysis was undertaken in order to predict utility values for all states described by the new health state classification.

Results: The proposed novel methodology resulted in the development of CORE-6D, a 2-dimensional health state classification consisting of a unidimensional 5-item emotional component and a physical item. Inspection of the Rasch item threshold map of the emotional component helped identify a set of 11 plausible emotional health states that are frequently observed and cover the full range of symptom severity in the study population, that is, in people with common mental health problems. These 11 emotional health states combined with the 3 response levels of the physical item of CORE-6D generate 33 plausible health states, 18 of which were selected for valuation. A number of multivariate regression models at the mean (aggregate) level were used to analyse the results of the valuation survey in order to predict values for all health states defined by CORE-6D, using the Rasch logit value of the emotional state and the response level of the physical item as independent variables. A cubic model with high predictive value (adjusted R^2 0.990) was selected to predict utility values for all 729 CORE-6D health states, resulting in the development of a new PBM for people with common mental health problems.

4. Evaluation of the performance of the new PBM relative to generic ones, the degree of loss of information relative to the original CSM, and its relevance to people with mental health problems.

Methods: A series of psychometric tests, statistical analyses and qualitative assessments were performed in order to evaluate the performance and the properties of CORE-6D. CORE-6D was compared with the generic PBMs EQ-5D and SF-6D and with the original CORE-OM in three datasets: a dataset derived from participants in a British national psychiatric morbidity survey; a dataset of people with mild to moderate anxiety and/or depression participating in a randomised controlled trial (RCT) assessing self-directed psychological therapy; and a dataset of postnatal women recruited for a multicentre RCT that evaluated psychological interventions for postnatal depression. CORE-6D was compared with generic PBMs in terms of acceptability, floor and ceiling effects, responsiveness to change over time, and construct validity (known groups and convergent); agreement between CORE-6D and generic PBMs and differences in the content of their items were also assessed. Furthermore, the content validity of CORE-6D and generic PBMs in the area of mental health was assessed by comparing the items of each measure against the 7 themes of HRQoL that have been found to be most important in people with mental disorders. The degree of loss of information was assessed by comparing the responsiveness and known groups validity of CORE-6D with those of CORE-OM.

Results: CORE-6D was shown to have comparable acceptability with generic PBMs and no floor or any significant ceiling effects. Results of analyses that tested its construct validity and responsiveness relative to generic PBMs were promising. CORE-6D shows acceptable agreement with generic PBMs and, in contrast to them, it broadly covers all 7 major themes of HRQoL that have been found to be important to people with mental disorders, although it is unable to capture some important sub-themes such as subjective well-being. CORE-6D showed small loss of information relative to the CORE-OM. Analyses suffer from a number of limitations that should be addressed in future research.

Discussion and conclusion: The CORE-6D is a promising PBM that appears to be relevant, valid and responsive in people with common mental health problems. CORE-6D will enable economic evaluation of mental health interventions in the form of cost-utility analysis, using existing and prospective CORE-OM datasets. The use of condition-specific PBMs instead of generic ones in the wider resource allocation context has raised concerns relating to their narrow scope and their inability to capture side effects from treatment or comorbidities, the distortions created in the valuation process by focusing effects, and potential bias in valuation resulting from naming the condition. Nevertheless, CORE-6D addresses some of the expressed concerns by covering physical problems to some extent. The novel methodology proposed in this thesis for the derivation of CORE-6D from the CORE-OM, including the role of Rasch analysis in the development of the health state classification, the selection of plausible health states and the subsequent modelling of utility values may be useful for similar processes using other instruments with highly correlated dimensions.

Areas for further research:

- To further explore the role of Rasch analysis in the development and valuation of health state classifications
- To conduct larger valuation surveys and explore the preferences of people with common mental health problems for HRQoL levels described by CORE-6D
- To further test and validate the applicability and performance of CORE-6D in other mental health conditions
- To map CORE-10 and CORE-5 onto CORE-6D so as to allow cost-utility analysis in studies that use CORE-10 or CORE-5 but not CORE-OM
- To explore the potential use of CORE-6D as an independent measure, when CORE-OM is not used in a study
- To expand the use of CORE-6D outside the UK
- To develop *de novo* a generic PBM measure for mental disorders

Chapter 1. Introduction and background

1.1 Introduction

Economic evaluation of healthcare technologies aims at optimal allocation of healthcare resources in order to maximise the health of the population. Resources are finite and thus choices on how to spend them in the most efficient way need to be constantly made. Formal economic evaluation allows a consistent, standardised way for making such choices, aiming at achieving the best possible overall health status for the population. Cost-utility analysis is a type of economic evaluation in which health outcomes are expressed in the form of a generic summary measure that combines length of life with *preferences* for different states of health experienced through life, on a single scale. The most commonly used outcome measure in cost-utility analysis is the Quality Adjusted Life Year (QALY), which expresses a person's life expectancy weighted by the 'utility value' of the health-related quality of life (HRQoL) experienced in each period of life. Utility values reflect people's preferences for HRQoL, measured on a scale anchored between zero (death) and one (full health), with negative values being attached to health states deemed worse than death (Brazier et al., 2007). When the QALY is used as the measure of outcome, then the output of cost-utility analysis is an incremental cost effectiveness ratio (ICER) expressed as 'cost per QALY gained'.

Generic preference-based measures (PBMs) such as the EQ-5D (Dolan, 1997), the SF-6D (Brazier et al., 2002) and the HUI-3 (Feeny et al., 2002) are widely used for the estimation of QALYs. However, their relevance in some disease areas, including mental health, has been questioned (Brazier, 2010; Brazier & Tsuchiya, 2010). This thesis is concerned with the development of a PBM that is relevant and sensitive to people with mental health problems. This chapter begins with an overview of the concept of mental health and its impact on people's lives and briefly discusses some issues on outcome measurement in mental health research and practice, focusing on the use, categories and methods of assessment of patient-reported outcome measures (PROMs). It then moves on to give a summary of the steps required for the development of PBMs which are used for outcome measurement in cost-utility analysis, a brief

description of the three most commonly used generic PBMs, and a note on the role of condition-specific measures (CSMs) in the estimation of QALYs. Subsequently it sets the context in which this thesis was undertaken, in terms of the burden of mental disease in the UK and the current levels of research in this area; after a short discussion of the concerns that have been raised regarding the suitability of generic PBMs in mental health populations, it provides the rationale for and the objective of this thesis. Finally, it presents an overview of the key stages of research undertaken to fulfil the main objective of the thesis, along with an outline of this thesis report.

1.2 Defining and measuring mental health – the role of patient-reported outcome measures

1.2.1 Mental health and its importance in people's well-being

According to the World Health Organization (WHO), health is defined as a "*state of complete physical, mental, and social well-being and not merely the absence of disease and infirmity*" (World Health Organization, 1958). This definition emphasises the importance of mental health as an integral component of health and well-being. Mental health is described as "*a state of well-being in which every individual realizes his or her own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to her or his community*" (World Health Organization, 2010).

Mental illness is one of the leading causes of disability globally, especially in high-income countries. Depression alone accounts for 4.3% of the global burden of disease and is among the largest single causes of disability worldwide (World Health Organization, 2008). People with mental disorders experience disproportionately higher rates of disability and mortality owing to physical health problems that are often left unattended (such as cancer, diabetes, cardiovascular disease and HIV infection) and suicide. Suicide is the second most common cause of death among young people worldwide. Mental disorders often affect, and are affected by, other diseases such as cancer,

cardiovascular disease and HIV infection / AIDS (World Health Organization, 2013).

Recently, WHO developed a mental health action plan, with the vision of "*a world in which mental health is valued, promoted, and protected, mental disorders are prevented and persons affected by these disorders are able to exercise the full range of human rights and to access high-quality, culturally appropriate health and social care* […] *in order to attain the highest possible level of health*" (World Health Organization, 2013). In order to achieve this, WHO set forth four major objectives: more effective leadership and governance for mental health; the provision of comprehensive, integrated mental health and social care services in community-based settings; implementation of strategies for promotion and prevention; and strengthened information systems, evidence and research. The last objective requires, among other interventions and strategies, collecting information on indicators of mental health, including data on suicide and premature mortality, as well as improvements related to clinical symptoms, levels of disability, overall functioning and quality of life (World Health Organization, 2013).

Looking at WHO's mental health action plan, it is evident that measurement of clinical symptoms, disability, functioning and HRQoL is considered to be important for the promotion of mental health and the prevention and treatment of mental illness.

1.2.2 Outcome measurement in mental health

The purpose of outcome measurement in medical research and practice is to evaluate the relative effectiveness and cost-effectiveness between treatments, interventions and technologies, to inform individual clinical decisions in routine practice by identifying and monitoring clinical conditions, to enable clinical audit, and to assess and monitor the health needs of a population (Gilbody et al., 2003). In the area of mental health, outcome measurement involves the verification and quantification of a wide range of psychiatric phenomena that cannot be externally observed or verified and are subjective in their nature (Gilbody et al., 2003). A broad category of instruments developed for diagnosing mental disorders, identifying and also quantifying mental symptom

severity comprises standardised, symptom-based psychopathology rating scales; these measures are usually clinician-rated and are routinely used in mental health research and practice (Gilbody et al., 2003).

Despite their value in assessing the existence and severity of symptoms, clinician-rated measures cannot assess the full impact of disease on patients' lives. In response to this need, PROMs have emerged in healthcare research and practice, including mental healthcare. These measures are completed by patients themselves rather than by clinicians or carers on their behalf, in an attempt to "*move the focus of healthcare evaluation from 'technical' outcome measures assessed by health professionals to those outcomes and aspects of health valued by the recipients of care*" (Jenkinson & McGee, 1998), giving the latter the opportunity to be involved in judgements regarding the effectiveness of services (Fitzpatrick et al., 2007). There is evidence that patients may consider distress associated with the effects of disease on their daily activity to be more disturbing than the discomfort associated with their symptoms (Janson-Bjerklie et al., 1992), indicating that the perceived burden of disease, expressed in the form of a disability or handicap, is possibly more important than the symptoms of the disease in determining the patients' HRQoL (Bowling, 1997).

The common element of PROMs or 'HRQoL' measures, or 'health status' measures (terms that have often been used interchangeably in the literature) is an attempt to directly capture patients' subjective perceptions and experiences of important aspects of their health status, including their physical, psychological and social functioning and well-being (Fitzpatrick et al., 2007; Leidy et al., 1999; McDowell & Newell, 1996). More specifically, HRQoL measures designed for use in people with mental disorders "*cover patients' perspectives on what they have, how they are doing and how they feel about their life circumstances*"; such perspectives include the sense of wellbeing, functional status, access to resources and opportunities (Lehman & Lasalvia, 2010). PROMs usually take the form of questionnaires with several items that cover the broad nature of health status, disease or injury; usually each item

gets a score depending on the response level, and item scores are summed to give a total score (Fitzpatrick et al., 2007).

PROMs are increasingly used for a broad range of purposes, including research, routine patient monitoring, population studies, audit and quality assurance, as well as at resource allocation decision-making (Fitzpatrick et al., 2007; Jenkinson & McGee, 1998). Advisory and regulatory bodies worldwide have published guidance on the use of PROMs. The Department of Health in the UK has introduced a PROMS programme in England, aiming to assess the quality of care provided by the National Health Service (NHS) from the patient perspective (Department of Health, 2008). Mental health services in the UK also include the routine assessment of PROMs in psychological services, via the Improved Access to Psychological Therapies (IAPT) initiative (IAPT, 2011). In the US, the Food and Drug Administration, that is, the consumer protection agency of the US Government that monitors the safety and effectiveness of human and veterinary drugs among other products, has published guidance on the use of PROMs in supporting labelling claims for approved medical products (US Department of Health and Human Services, Food and Drug Administration, 2009).

1.2.3 Generic and specific patient-reported outcome measures

PROMs are divided into two broad categories: generic and specific (Fitzpatrick et al., 2007). Generic measures aim to cover multiple domains of HRQoL applicable to a wide range of disease areas and populations, and therefore allow comparisons of treatments provided to different patient groups with a variety of conditions, thus enabling assessment of comparative effectiveness (Fitzpatrick et al., 1998). An advantage of generic measures is that they can be useful in the identification of comorbidities and side-effects of treatment that cannot be captured by specific instruments (Fitzpatrick et al., 2007). This property makes them suitable for assessing the impact of new healthcare interventions, where the therapeutic effects may be uncertain or even unknown (Cox et al., 1992; Fletcher, 1988). On the other hand, generic measures are designed to capture broad aspects of HRQoL and therefore may be less relevant to specific conditions, as it is not possible to capture all particular

dimensions of HRQoL relating to every disease area. Consequently, in such cases they may have lower responsiveness to small but important changes in HRQoL (Fitzpatrick et al., 1998 & 2007). The most commonly used generic measures are the EQ-5D (Brooks, 1996), the Short Form (36) Health Survey (SF-36) (Ware et al., 1993) and its shorter forms SF-20 (Ware et al., 1992) and SF-12 (Ware et al., 1995), and the Health Utilities Index [HUI] system (Torrance et al., 1995).

In contrast to generic measures, specific instruments concentrate on a particular disease area, patient population, symptom, function, aspect of HRQoL, or part of the body and are therefore very relevant to their intended focused field (Fitzpatrick et al., 1998 & 2007). Thus they have been argued to show high sensitivity in detecting small but significant changes in HRQoL observed in the area of interest (Wiebe et al., 2003). Due to their relevance to a particular situation, specific instruments are deemed to be more acceptable to the respective patient population (Fitzpatrick et al., 1998). Their main disadvantage is that they do not allow comparisons of treatment outcomes across patients with different health problems. They also lack the ability to capture symptoms and side effects of treatment that are unusual or unexpected in the area which they have been specifically designed for (Fitzpatrick et al., 1998 & 2007).

1.2.4 Criteria for evaluating patient-reported outcome measures

Classical psychometric tests are widely used for the assessment of PROMS. These tests are applications of standard psychometric criteria against which an instrument should be examined. Although all of these criteria are important in evaluating PROMs, some of these are more frequently listed on respective assessment checklists (for example, Mokkink et al., 2010; Terwee et al., 2007). Such criteria are also relevant and have been used in assessing measures utilised in economic evaluation (Brazier et al., 1999). Some of these criteria can also be used at the selection of ‘best performing’ items for the derivation of a *health state classification* system from a larger questionnaire, an application that is discussed in Chapter 3 (section 3.3.1). The psychometric criteria and

the respective psychometric tests available for the evaluation of PROMs are (Fitzpatrick et al., 1998):

Appropriateness

The content of an outcome measure needs to be appropriate to the aims of a particular study. Appropriateness is difficult to evaluate and mainly relies on researchers' judgement as to whether the content of an instrument (and its individual items) is in line with the study questions, after taking into consideration the nature of the study intervention, the study population, and the health outcomes to be captured. Although appropriateness is an essential characteristic of a PROM, no standard psychometric tests are available to directly evaluate this criterion, and other psychometric criteria of those listed below need to be examined to indirectly assess the appropriateness of a measure (Fitzpatrick et al., 1998).

Reliability

Reliability is the ability of an instrument to provide measurement results free from random error. Reliability is directly related to the internal consistency of a measure and its reproducibility.

Internal consistency refers to the homogeneity of a measure, that is, "*the extent to which all its items measure aspects of a single attribute or construct*" (Streiner & Norman, 1995). Consequently, all individual items of an internally consistent measure are expected to be highly correlated with each other and with the total score of the measure (summed score of its items). There are two main psychometric tests used to assess internal consistency: one approach is to randomly divide the items of an instrument into two groups and to assess the degree of agreement (correlation) between the two halves (split-half reliability). An extension to this approach is the estimation of Coefficient alpha (Cronbach's α), which estimates the average level of agreement of all possible ways of performing split-half tests (Cronbach, 1951). Cronbach's α needs to be high enough to ensure sufficient internal validity of a measure; on the other hand, perfect correlation of items indicates that these capture a rather narrow

aspect of an attribute. For this reason, a value of Cronbach's α between 0.70 and 0.90 has been suggested as optimal (Streiner & Norman, 1995).

Reproducibility is the ability of a measure to reproduce the same results if it is repeatedly administered to the same population, when the latter has not changed with respect to the characteristics being measured. A measure with perfect reproducibility should be able to provide results free from random error. Reproducibility is assessed by test-retest reliability, which examines the agreement between scores of two assessments on the same study sample between two different time points, during which the study sample is unlikely to remember their previous responses but at the same time it is also unlikely to have changed with regards to the health dimension assessed. Test-retest reliability is evaluated by a correlation coefficient, such as the intra-class correlation coefficient, which uses analysis of variance to determine the extent of total variability between the two scores that is due to true differences between respondents and due to variability in measurements (Fitzpatrick et al., 1998).

Validity

Validity refers to the extent to which an instrument measures what it is intended to measure. There are several ways to assess the validity of an instrument, such as face validity, content validity, construct validity, criterion validity and predictive validity. The first three aspects of validity are by far the most relevant and widely used criteria for the assessment of PROMs (Fitzpatrick et al., 1998).

Face validity examines whether an instrument "*appears to be measuring what it is intended to measure*" (Guyatt et al., 1993). This assessment relies mainly on judgement of the content of an instrument.

Content validity examines the extent to which the characteristic of interest (e.g. HRQoL) is comprehensively captured by the items of the instrument (Guyatt et al., 1993). As with face validity, content validity cannot be directly assessed by

psychometric tests and therefore judgement (by either researchers or patients) is required to estimate whether a measure is characterised by this attribute.

Construct validity refers to the extent to which an instrument is able to measure the underlying 'construct' which it was designed to assess. Construct validity can be assessed by the ability of the measure to distinguish between groups that are known to differ in the underlying construct ('known groups validity'), or by examining the correlations of the instrument with a set of other variables that have been designed to measure the same ('convergent validity') or a different ('discriminant validity') construct (Fitzpatrick et al., 1998 & 2007).

Criterion validity assesses whether an instrument correlates with another instrument that is accepted as an accurate measure of the attribute under measurement. Ideally the instrument should be compared against a 'gold-standard' measure ('*criterion variable*'), which, nevertheless, rarely exists in the area of health status measurement (Fitzpatrick et al., 1998).

Predictive validity is an attribute relating to the ability of the measure to correlate with future values of the criterion variable (Fitzpatrick et al., 1998).

Responsiveness

Responsiveness expresses the ability of an instrument to "*detect important changes within individuals that may reflect therapeutic effects*" (Kirshner & Guyatt, 1985). Responsiveness of an instrument can be assessed by employment of various statistical methods, such as (Fitzpatrick et al., 1998):

- Correlation of the change score of an instrument over time with change scores of other measures that are intended to capture the same changes
- Effect size (ES), defined as the change score of a measure divided by the standard deviation at baseline
- Standardised response mean (SRM), defined as the change score of a measure divided by the standard deviation of the change score

The responsiveness of a measure can be affected by factors ‘internal’ to the content of a questionnaire. One example is the presence of ‘floor’ or ‘ceiling’ effects. These may exist when the instrument is not well targeted to the study population, in the sense that it cannot measure the whole range of severity of the health dimension in question. Therefore, a measure may be unable to capture significant improvement or deterioration if the initial or final point of change lies beyond the range of severity the instrument is able to capture (Fitzpatrick et al., 1998).

Precision

Precision of an instrument relates to its ability to distinguish among different levels and aspects of the health dimension under assessment. There are various aspects related to the precision of an outcome measure (Fitzpatrick et al., 1998):

Precision of response categories: Precision of an instrument is improved when item responses are graded across multiple rather binary response categories, as in this case it is possible to capture more accurately the various levels of the health dimension under assessment.

Precision of numerical values: This type of precision relates to the ability of numerical values of a PROM in accurately capturing subjective experiences. Numerical values can take two forms: simple ordinal values, where for example degrees of agreement with a statement are attached a progressively increasing or decreasing value; and weighted numerical values, which are given to individual items following judgements on their relative severity/ importance.

Even distribution of item responses over true range: a PROM needs to capture the full range of the severity of the examined dimension and distinguish well across intermediate severity levels, that is, it needs to be well targeted to the patient population. Inability to capture the full range of symptoms can affect the responsiveness of an instrument, as discussed above.

Floor and ceiling effects: these are related to the distribution of items over the full range of severity of the health dimension examined, and their presence affects negatively the precision of an instrument in capturing differences across the levels of the health dimension. Floor and ceiling effects indicate a high proportion of responses at the lower and higher end of a scale, respectively.

Precision of scales: this type of precision relates to whether an instrument measures the construct which it is aimed to assess rather than other unrelated aspects. Ideally, for this purpose, the scales of an instrument should be unidimensional. One method to assess the dimensionality of an instrument is by factor analysis, which is described in detail in Chapter 3 (section 3.3.1).

Interpretability

PROMs may be more difficult to interpret compared to physical/ clinical measures such as, for example, blood pressure or blood sugar levels. Interpretability of a PROM is increased if the measure's scores are correlated to other human experiences with a clear and more objective meaning, such as side effects from treatment and stressful life events (Testa et al., 1993). Another approach suggested for increasing the interpretability of PROM scores is to determine a minimal clinically important difference, defined as "*the smallest difference in score which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive costs, the patient's management*" (Jaeschke et al., 1989). Comparison of PROM scores between a clinical population and the general population can also enhance interpretability of the results (Fitzpatrick et al., 1998).

Acceptability

Acceptability to patients is an important attribute of a PROM. Acceptability of a measure relates to the distress potentially caused to patients completing a questionnaire. An instrument may be less acceptable if it is difficult to understand or requires a long time to complete. Acceptability of an instrument can be indirectly assessed by estimation of completion rates following administration; acceptability of certain items of an instrument can be examined

by measuring the proportion of missing data in completed questionnaires. Time to complete an instrument may also be an indicator of patients' acceptability for a measure (Fitzpatrick et al., 1998).

Feasibility

Feasibility refers to the "*time and resources required to collect, process and analyse a patient-based outcome measure*" (Fitzpatrick et al., 1998). The length and complexity of a questionnaire are factors affecting its feasibility.

1.3 Outcome measurement in cost-utility analysis

PBMs comprise a special type of PROM developed to provide an estimate of patients' preferences for different levels and aspects of HRQoL (Fitzpatrick et al., 1998). They consist of a health state descriptive system describing aspects of HRQoL, and a scoring algorithm converting the HRQoL in each health state into a utility value, based on preferences elicited in a valuation survey. As noted in section 1.1, PBMs are essential for the estimation of QALYs in cost-utility analysis.

1.3.1 Steps in the estimation of utility values

Estimation of preferences for different health states is a 3-step process, consisting of identification and *description* of the health states characterising a disease area, population or condition; *valuation* of a selection of those health state descriptions; and, finally, application of *modelling* techniques to valuation data in order to attach preferences to all relevant health states described by a health state descriptive system.

Description of health states

Health states can be described using, mainly, a generic or condition-specific outcome measure. A health state is constructed by selecting one response level from each item of the measure, and combining all item responses. Outcome measures that are used for the description of health states comprise *health state classifications*. In a health state classification each item typically represents a separate dimension. Health states used for measurement and valuation of HRQoL should focus on functional status (physical, emotional and social) rather than clinical characteristics or laboratory test results (Torrance,

1986) and should therefore be derived from PROMs rather than clinician-rated outcome measures.

Health states can also be described by vignettes; the latter are usually narrative descriptions constructed based on interviews with patients and clinical experts, capturing various aspects of HRQoL, such as clinical symptoms, level of physical and social functioning, treatment and side effects (Brazier et al., 2007). Vignettes are in principle condition- and treatment-specific. Finally, sometimes health states are not described, but instead patients are asked to value their own health.

The characteristics, advantages and disadvantages of generic and condition-specific HRQoL measures have been briefly described in section 1.2.3 of this chapter and are discussed in detail in Chapter 8 (section 8.4.1). The drawbacks of vignettes include their inability to describe the full range of health states that are usually observed in a patient population, as each vignette describes only one state. Therefore, although vignettes can provide detailed descriptions of specific health states, they often lack the sensitivity to capture small variations in HRQoL. Moreover, vignettes may be difficult to link to outcomes reported in clinical trials. Finally, the psychometric properties of vignettes are more difficult to empirically assess compared with standardised measures (Brazier et al., 2007).

Valuation of health states

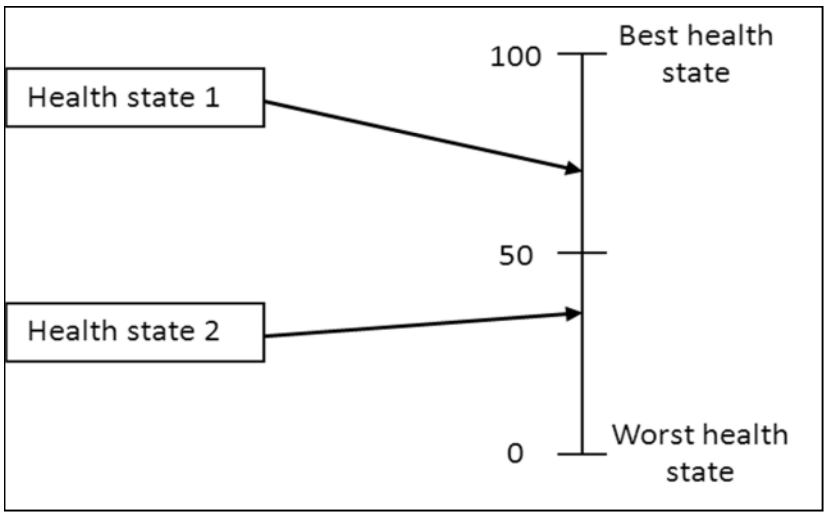
Valuation of a health state refers to attaching a *preference* to the HRQoL represented by the state. Preferences may be elicited by patients, their carers, health professionals, or members of the general public. Until recently, there were 3 main methods for preference elicitation: the Visual Analogue Scale (VAS); the Standard Gamble technique (SG); and the Time Trade-Off technique (TTO). More recently, there is an increasing interest for valuation of health states by methods that are based on collection of ordinal information, such as ranking and Discrete Choice Experiments (DCEs) (Ali & Ronaldson, 2012; Brazier et al., 2007). It has been argued that preferences express ‘values’ if the framing of the question does not involve uncertainty, as in the

VAS and TTO, and ‘utilities’ if the question requires consideration of uncertainty, as in the SG (Drummond et al., 2005). In this thesis, the term ‘utility value’ will be used more generally to describe preferences elicited from valuation studies, regardless of the method of valuation and the framing of the question.

Visual Analogue Scale

The VAS is a simple line with defined anchor states, such as ‘full health’ or ‘best health state’ on the one side of the line and ‘death’ or ‘worst possible health state’ on the other. Respondents are asked to place their preference for specific health states along the line. The scale has interval properties, so that the distances between the placements of health states correspond to the respondents’ relative differences in preference between the states (Neumann et al., 2000; Torrance, 1986). Anchoring the scale between ‘best imaginable state’ and ‘worst imaginable state’ allows valuation of the ‘death’ state and elicitation of preferences for states considered worse than death (Feeny et al., 2002; Gudex et al., 1996). It has been argued that using clear and unambiguous endpoints on the scale ensures comparability of judgements between respondents (Brazier et al., 2007). A simple graphic example of VAS is shown in Figure 1.

Figure 1. Illustration of a Visual Analogue Scale

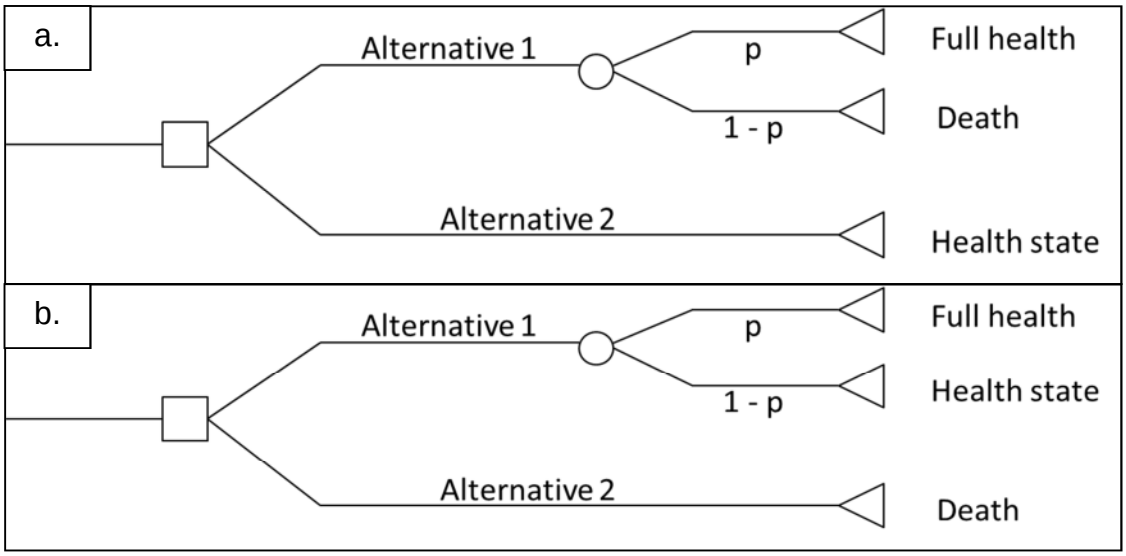


Standard Gamble

The SG technique asks respondents to consider the level of risk which they are willing to take with their life in a certain health state in order to return to full health. It is based on the axioms of the von Neumann-Morgenstern utility theory, according to which when rational individuals are faced with a choice between options they will choose the option that maximises their expected value of utility (von Neumann & Morgenstern, 1953). SG gives the respondent a choice between a certain intermediate outcome and the uncertainty of a gamble between two possible situations, one of which is better than the certain outcome and the other is worse. For chronic health states considered better than death, alternative 1 involves a gamble between life in full health for *t* years (probability p) or immediate death (probability 1 – p); alternative 2 is the certain outcome of life in the health state for *t* years. The probability p is varied until the respondent is indifferent between the gamble and the certain outcome. At this point the probability p expresses the utility value attached on the health state. For chronic health states considered worse than death, alternative 1 involves a gamble between life in full health for t years (probability p) or life in the health state in question for *t* years (probability 1 – p); alternative 2 involves the certain outcome of immediate death. The probability p is varied until the respondent is indifferent between the gamble and the certain outcome. The utility value of a health state deemed worse than death is then given by the formula –p/(1-p) (Brazier et al., 2007; Torrance, 1986).

A schematic diagram of the SG task for chronic health states is provided in Figure 2.

Figure 2. Schematic diagram of standard gamble for a chronic health state a) preferred to death and b) considered worse than death



Time Trade-Off

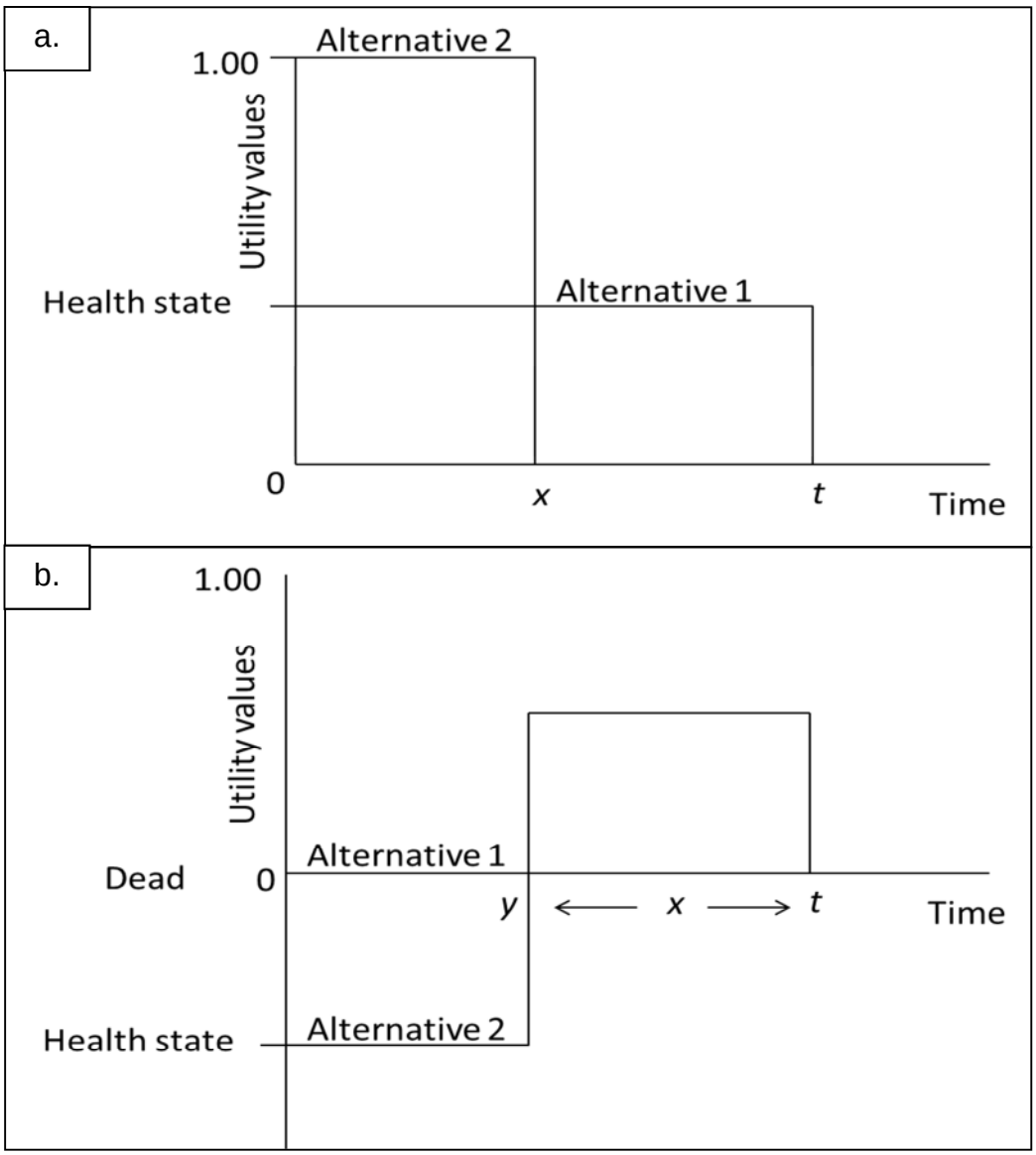
The TTO technique was suggested by Torrance and colleagues (1972) as an alternative to SG that is simpler to use but provides similar results. Unlike SG, TTO elicits decisions under certainty. The TTO task asks respondents to trade HRQoL for life-years. More specifically, for a specified health state (h_i) that is worse than full health but better than death respondents are asked to choose either to live for a period of t years in this state, or to shorten their lifespan to x years in full health, where $x < t$. The number of x years in full health is varied, until the point where the respondent is indifferent or switches preferences between the two alternatives. The utility value given to the state h_i is then x/t (Brazier et al., 2007).

For health states considered worse than death, the TTO task can be modified. For example, in the Measurement and Valuation of Health study (MVH) (MVH Group, 1995) that was used at the valuation of EQ-5D (Dolan et al., 1996), respondents were first asked whether they preferred to live in a specified health state h_i for a period of t years after which they would die, or die immediately. This question determined whether respondents valued the health

state as better, worse, or equal to being dead. Subsequently, for health states considered worse than death, respondents were asked to choose between two alternatives: alternative 1 involved immediate death, while alternative 2 involved life in the health state for *y* years followed by life in full health for *x* years (with $y + x = t$) followed by death. Years in full health (*x*) were varied concurrently with years in the health state (*y*) so that *t* remained constant, until respondents were indifferent between the two alternatives. The utility value given to the health state in this case is $-x/y$ (Brazier et al., 2007; Torrance, 1986). However, this formula may produce very low values (in the case of MVH TTO protocol where $t = 10$ the lowest possible value for a state worse than death can reach -39), which creates problems when modelling valuation data, as values corresponding to states worse than death have a larger impact on the model predictions than values of states better than death (Rowen & Brazier, 2011). It has been therefore suggested that utility values for states considered worse than death be rescaled, so that they are bounded by a value of -1 (Torrance, 1984), and this approach was followed at the valuation of EQ-5D, where utility values of states worse than death were calculated by the formula $-x/t$ (Dolan, 1997; Dolan et al., 1996).

A schematic diagram of the TTO task for chronic health states is shown in Figure 3.

Figure 3. Schematic diagram of time trade-off for a chronic health state a) preferred to death and b) worse than death



Comparison between main valuation methods

The three main techniques for valuing health states (VAS, SG and TTO) have demonstrated satisfactory reliability and high acceptability to respondents (Green et al., 2000). However, they have been shown to result in different sets of values for the same health state descriptions (Hornberger et al., 1992; Read et al., 1984; Torrance, 1976). Various arguments in favour of or against the use of each of them have been expressed in the published literature:

The VAS method appears to be the easiest to comprehend and most acceptable to respondents (Drummond et al., 2005; Green et al., 2000). However, as the method does not require respondents to make a choice by trading-off different arguments in their utility function, it has been regarded as theoretically inferior to the choice-based TTO and SG (Dolan, 2001). In addition, VAS is subject to measurement bias, as elicited scores often lack interval properties (Drummond et al., 2005). This may explain why VAS values have only poor to moderate correlation with values derived from TTO and SG undertaken at the same time, while TTO and SG correlate reasonably well with each other (Green et al., 2000). Other problems characterising VAS include the 'end-of-scale' bias, meaning that respondents tend to avoid using the two ends of the scale, and 'spacing out' bias, in which respondents tend to spread their preferences on the scale, regardless of the nature of the health states (Drummond et al., 2005). Due to the types of bias inherent in the method, it has been argued that raw (unadjusted) VAS values cannot provide a valid basis for estimating preferences for health states (Brazier et al., 2007). Nevertheless, other researchers have argued that VAS valuations do involve choice-making, as respondents weigh-up pairs of health state descriptions with potentially small differences across the described HRQoL dimensions against the 'anchor' states and other health states included in valuation, and this decision actually involves making choices and trading-off between improvement in one HRQoL dimension and deterioration in another (Parkin & Devlin, 2006). Moreover, appropriate transformation of raw VAS data may remove bias resulting from 'end-of-scale' and 'spacing out' phenomena (Parkin & Devlin, 2006).

The SG has been advocated by economists because it entails making decisions under uncertainty, which also surrounds most decisions about health care (Mehrez & Gafni, 1993). Yet, it has been argued that the appropriateness of a valuation method should be determined by its ability to act as a proxy for utility and not by its capacity to model the situation being valued (Dolan, 2001). SG may be compromised by probability weighting, according to which respondents tend to overweight small probabilities and underweight large

ones; if the probability weighting function is inverse S shaped as indicated by empirical evidence, and the point where the function changes from overweighting to underweighting probabilities approximates 0.35 as suggested in the literature, then SG tends to overestimate utility values given that the probabilities reported in SG exercises overall tend to exceed 0.35 (Bleichrodt, 2002). Moreover, SG is also affected by risk aversion resulting in SG values being pushed upwards, and scale compatibility (this is where respondents assign more weight to attributes that have higher compatibility with the response scale used) resulting on respondents' focusing on the probability rather than the health state valued; and because there are more than one probabilities involved in the task, the direction of bias in estimation of utility values cannot be predetermined (Bleichrodt, 2002).

TTO has been considered the most appropriate valuation method, as it incorporates the relationship between the health state, its duration and its value into a single measure (Dolan, 2001). There is evidence, however, that TTO values are prone to duration and time preference effects; in other words, the period of time spent in a health state and the point in time a health state is experienced (e.g. at the beginning or end of a time period) affect the way the state is perceived by respondents and therefore have an impact on utility values (Dolan & Gudex, 1995). Moreover, TTO assumes that utility is linear in duration, and given that utility has been empirically shown to be concave, the TTO task tends to systematically underestimate utility values (Bleichrodt, 2002). The assumption of linearity is more strongly violated in end-of-life scenarios (Garau et al., 2011). Another issue is that TTO is affected by attitudes such as loss aversion (so that respondents tend to be more reluctant to give up healthy life-years), and scale compatibility (so that respondents place more weight on the duration of a health state, which is the response scale of the task, rather than to the health state itself); both phenomena result in an overestimation of utility values (Bleichrodt, 2002).

Currently, TTO and SG are the most widely used techniques for valuation of health states (Brazier et al., 2007). Nevertheless, VAS has often been used for respondents' warming-up prior to TTO and SG exercises, so as to familiarise

respondents with descriptions of health states and give them an opportunity to start considering their preferences (e.g. Brazier et al., 1998; Dolan et al., 1996).

Valuation methods based on ordinal data: Ranking and Discrete Choice Experiments

Valuation methods using ordinal data are increasingly used due to a number of advantages compared with the 'standard' TTO and SG techniques, including their ease of administration and comprehension by the respondents and the avoidance of responses being affected by risk aversion, time preference, and other biases characterising TTO and SG (Ali & Ronaldson, 2012; Brazier et al., 2007). In ranking, respondents are asked to order a number of health states from the best to the worst. In DCEs, respondents are asked to make stated choices and select one state between two alternatives or make choices amongst a larger set of alternative options. Questions may be framed as the state the respondent would select to live in for a defined period of time, or the state that corresponds to the best health level. Ordinal data can be subsequently transformed into cardinal data (utility values) using statistical methods such as logit and probit modelling (Ali & Ronaldson, 2012; Brazier et al., 2007; de Bekker-Grob et al., 2012). A key problem with this method is how to anchor the values on the 0-1 scale required to generate QALYs, where 1 is for full health and 0 for states as bad as being dead. There is ongoing research looking at the use of duration as an additional attribute (e.g. Bansback et al., 2012).

Modelling valuation data

The process of valuation cannot be applied to all potential health states described by an instrument, as this would be extremely time- and resource-consuming due to the high number of health states that can be described by one instrument. For example, EQ-5D can describe 243 different health states, while the number of health states that are described by SF-6D reaches 18,000. Instead, a number of health states described by an instrument are selected for valuation; subsequently, using the utility data obtained in the valuation survey, modelling techniques are employed to attach an appropriate utility value to every health state described by the measure. There are two main approaches

for modelling utility values: the composite approach, which uses statistical modelling to estimate an algorithm for valuing all health states described by an instrument using utility data derived from valuation of selected health states; and the decomposed approach, which employs Multi-Attribute Utility Theory (MAUT) to determine the functional form underlying the relationship between single dimensions as well as the sample of states to be valued (Brazier et al., 2007). A prerequisite for using any of the two approaches is the multidimensionality of the instrument to be valued, i.e. each item of the instrument needs to be independent from the rest items and express a different dimension of HRQoL.

The composite approach for modelling valuation data has been used in the valuation of the EQ-5D (Dolan, 1997), the SF-6D (Brazier et al., 2002), and several condition-specific PBMs (for example Brazier et al., 2008; Yang et al., 2009 & 2011). The first step of this approach relies on the identification and selection of a set of health states described by the instrument, in order to be included in a valuation survey. Selection of health states can be achieved using a statistical design such as an orthogonal array, which allows the statistical testing of several factors without testing every combination of factor levels (Hedayat et al., 1999). Alternatively, a balanced approach can be adopted, which allows any response level of each dimension to have an equal chance of being combined with the various response levels of all the other dimensions comprising the instrument (for example Yang et al., 2011). As the number of selected health states can be still quite large, the selected health states may be divided in smaller subsets that are valued by different groups of the valuation survey participants. Following the survey, a number of regression models are fitted to the valuation data, aiming to identify the model that best describes the relationship between the valued health states and the utility values obtained from the survey, which is then used to predict utility values for all states described by the instrument (Brazier et al., 2007). The model specifications can be quite complex, as they need to take into account the non-normality and the quite commonly observed skewness of the utility data, the non-continuity of the data distribution, and the fact that different states are

valued by different survey participants, so data are also likely to reflect differences in participants' preferences (Brazier et al., 2007).

A general model used to predict utility values for all potential health states of an instrument using valuation survey utility data was described for the statistical modelling of utility values of SF-6D (Brazier et al., 2002), and is defined as follows:

$$y_{ij} = g\,(\beta' x_i + \theta' r_i + \delta' z_j) + \varepsilon_{ij}$$

where $i = 1, 2, \ldots, n$ represents individual health state values and $j = 1, 2, \ldots, m$ represents respondents. The dependent variable, y_{ij} is the adjusted utility value for health state i valued by respondent j, x is a vector of binary dummy explanatory variables ($x_{\delta\lambda}$) for each level λ of dimension δ of the instrument; r is a vector of terms to account for interactions between the levels of different dimensions and z is a vector of personal characteristics such as age, gender and education, which may affect values placed by an individual on a health state; g is a function specifying the appropriate functional form and ε_{ij} is an error term whose autocorrelation structure and distributional properties depend on the assumptions underlying the particular model used. This model specification represents a simple additive function, as it imposes no further restrictions on the relationship between dimension levels of the instrument (e.g. it does not impose an interval scale between the levels of each dimension) (Brazier et al., 2007).

Statistical modelling is possible to consider individual respondent data, or data at an aggregate (population) level (Brazier et al., 2002 & 2007). Models analysing individual respondent data can take into account the impact of respondent background characteristics, such as gender, age, socioeconomic status, etc., on health state valuations. The ordinary least squares (OLS) model, which allows prediction of utility values by linear regression, is the simplest model that can be used for the analysis of individual respondent data; this model specification ignores the between-respondent variation and assumes that each individual utility value is an independent observation, regardless of which respondent it was elicited from (Brazier et al., 2007). A

more sophisticated model, which takes into account the variation in valuation data both within and between respondents, is the one-way error components random effects model; this model assumes that the error is distributed between respondent-specific variation, and an error term for every health state valued by each respondent, both of which are assumed to be random across individual respondents. Such a specification can be estimated using a generalised least squares or a maximum likelihood model (Brazier et al., 2007).

The aggregate model ignores individual respondent characteristics and instead analyses population-level (mean or median) utility values; such specification is also estimated by an OLS model. Although statistical modelling that considers individual respondent data would be expected to predict utility values more accurately since it increases the number of degrees of freedom available for analysis (Brazier et al., 2007), valuation of SF-6D showed that this is not necessarily the case, as OLS aggregate models were shown to perform better than individual ones (Brazier et al., 2002). Since then, other modelling studies have replicated this finding (Brazier et al., 2008; McKenna et al., 2008; Yang et al., 2009).

The decomposed approach is based on MAUT, which has been mainly used at the development of the HUI utility system (Feeny et al., 2002; Torrance et al., 1996). MAUT uses simplifying assumptions about the underlying relationship between dimensions, determining how dimensions and dimension levels can interact with each other; the most commonly used specifications are the additive, the multiplicative and the multi-linear functional forms. The additive functional form assumes that dimensions are independent, and does not allow for any interactions between them; it simply adds up the utility ‘decrements’ associated with loss of HRQoL within each dimension. The multiplicative function permits limited interaction between dimensions, by assuming preference dependence to be the same between dimensions. When the combined decrement between any two dimensions is assumed to exceed the sum of the individual effects of the two dimensions, then dimensions are substitutes; when the combined decrement between any two dimensions is

assumed to be lower than the sum of the individual effects of the two dimensions, then the dimensions are complements. The multi-linear function is the least restrictive among the MAUT functional forms, as it allows interactions between pairs of states, as well as higher order interactions, to be estimated independently, without imposing any restrictions on the direction of the preference dependence. Its drawback is that it requires a substantial amount of valuation work in order to be parameterised (Brazier et al., 2007). The application of MAUT involves three steps in the valuation process: first, every dimension of a measure is valued separately, assuming that all other dimensions are at the best response level, so as to obtain single-attribute utility values; next, 'corner' multidimensional states are valued, which consist of one dimension at one extreme (usually the worst response level) and the remaining dimensions at the other extreme (usually the best response level), requiring that dimensions are independent from each other in order to create meaningful health states; finally, a set of multidimensional states are valued, the choice of which is determined by the selected model specification. A simple additive model requires valuation of two multidimensional states only; an extra state is required when a multiplicative model is used, to allow estimation of the interaction between the states. Following valuation, prediction of utility values for all potential health states described by the measure can be achieved by solving a system of equations that allows calculations of utility decrements for every dimension and every parameter that reflects the preference interactions specified in the model (Brazier et al., 2007).

Valuation of health states and subsequent modelling of utility values for every possible health state described by a health state classification results in the development of a PBM that allows not only measuring but also valuing the HRQoL associated with a health condition, according to expressed preferences.

1.3.2 Generic preference-based measures

Generic PBMs, such as the EQ-5D (Brooks, 1996), the SF-6D (Brazier et al., 2002), and the HUI-3 (Feeny et al., 2002), are most widely used for the estimation of QALYs in cost-utility analysis. A brief description of these measures follows.

EQ-5D

The EQ-5D is a 5-item instrument capturing 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression; in the original structure of EQ-5D, each item has 3 response levels, ranging from no problems to extreme problems or inability to perform a task (Brooks, 1996). The EQ-5D health state classification can thus describe $3^5 = 243$ unique health states. A number of health states have been valued by a representative sample of the general population in the UK (as well as in a number of other countries) using VAS (Gudex et al., 1996) and TTO (Dolan et al., 1996). Further econometric modelling has resulted in the development of an algorithm that links each health state described by EQ-5D with an appropriate utility value, thus allowing the use of EQ-5D in cost-utility analysis (Dolan, 1997). More recently, a 5-level response version of EQ-5D was developed by the EuroQol Group (Herdman et al., 2011) with valuation being under way.

SF-6D

The SF-6D can be derived from the SF-36 (Brazier et al., 1998 & 2002) as well as from its shortened version SF-12 (Brazier & Roberts, 2004). It has 6 dimensions (physical functioning, role limitations, social functioning, bodily pain, mental health, vitality). When derived from the SF-36, physical functioning and bodily pain have 6 levels of severity, role limitations 4 and the rest 3 dimensions have 5 levels of severity each, which, combined, can describe 18,000 unique health states. When derived from the SF-12, SF-6D includes 3 levels of response for physical functioning, 4 levels of response for role limitations, and 5 levels of response for each of the remaining dimensions, resulting in the formation of 7,500 unique health states. A number of SF-6D health states have been valued by members of the general population in the UK (as well as in other countries) using SG (Brazier & Roberts, 2004; Brazier et al., 2002). Further econometric modelling has led to the development of an algorithm that can predict an appropriate utility value for each health state described by SF-6D.

HUI-3

HUI-3 belongs to a family of health state classification systems (Torrance et al., 1995) and has been designed specifically for use in adult populations. It covers 8 attributes, including cognition, vision, hearing, speech, ambulation, dexterity, emotion and pain (Feeny et al., 2002). Each attribute has 5 or 6 response levels. HUI-3 can describe 972,000 unique health states. Utility values for all states have been predicted using MAUT, following a valuation survey of members of the general population in Canada, which used VAS and SG.

The structure of the EQ-5D, SF-6D and HUI-3 health state classifications is presented in Appendix 1.

1.3.3 The role of condition-specific measures in the estimation of QALYs

Despite their widespread use, generic PBMs may be inappropriate or insensitive in capturing relevant aspects of HRQoL in some medical conditions (Brazier & Tsuchiya, 2010; Brazier et al., 1999), including hearing loss (Yang et al., 2013b), visual impairment due to macular degeneration (Espallargues et al., 2005), venous leg ulcers (Walters et al., 1999), urinary incontinence (Haywood et al., 2008) and overactive bladder (Kobelt et al., 1999), chronic obstructive pulmonary disease (Harper et al., 1997), and chronic schizophrenia (van de Willige et al., 2005). In such cases CSMs can be used instead, in order to capture more accurately and responsively changes in HRQoL.

However, the vast majority of the available CSMs have been developed in order to describe and *measure* rather than *value* HRQoL (that is, they are not preference-based), and therefore are not suitable for the calculation of QALYs in cost-utility analysis. There are two main approaches in order to overcome this problem and enable use of CSMs for estimation of QALYs: one approach is the “mapping” from CSMs directly onto generic PBMs (Brazier et al., 2010); the other approach is the development of preference-based CSMs either *de novo* (for example Revicki et al., 1998a & 1998b) or from existing CSMs (for

example Brazier et al., 2005, 2008 & 2010; Rowen et al., 2011; Sundaram et al., 2010; Yang et al., 2009 & 2011).

Mapping

Mapping from a non-PBM onto an existing PBM refers to the estimation of a relationship between the two measures, which can be made using expert opinion or empirically, using statistical association (Brazier et al., 2007). Mapping based on expert opinion relies on the judgements of professionals or researchers and has been criticised for its arbitrariness and, usually, for lack of any validity testing (Brazier et al., 2007). Empirical estimation of a mapping function between a non-PBM and an existing PBM is achieved by employing regression techniques, which are used in a dataset containing patient-level data on both the non-PBM and the PBM to determine a statistical relationship between the two measures. The estimated mapping function can be utilised in datasets that contain only the non-PBM, in order to indirectly estimate utility values derived from the PBM index (Brazier et al., 2007).

The mapping function can be determined using a simple additive model, where the total score of the non-PBM is regressed onto the PBM. The limitation of such a model is that it implicitly assumes that all dimensions of the non-PMB are equally important, all its items carry the same weight, and the item response levels have interval-scale properties (Brazier et al., 2007). More complex model specifications may take into account the dimension scores, item scores or item response levels of the non-PBM as independent variables and are possible to introduce interaction terms between dimensions and between items (Brazier et al., 2010). The drawback of such approaches is that they can result in a large number of independent variables, which, nevertheless, can be limited if, for example, items with non-significant coefficients are excluded. Another complex modelling approach is to estimate separate regression models between the non-PBM and each dimension of the PBM (Brazier et al., 2010).

Mapping suffers from a number of limitations, such as limited performance in terms of model fit (Tsuchiya et al., 2002) and inability to accurately predict

values across the spectrum of symptom severity (for example Gray et al., 2006). A major limitation inherent to the approach is that it assumes that the PBM covers all aspects of HRQoL captured by the non-PBM. However, where there is not sufficient overlap between the two measures the validity of the resulting mapping function is limited. For the above reasons, there has been an increased interest in the development of preference-based CSMs (Brazier et al., 2010).

Development of preference-based condition-specific measures

Development of a preference-based CSM requires a 3-step approach, consisting of the construction of a health state classification, valuation of a selection of health states, and employment of modelling techniques that allow prediction of utility values for all health states described by the classification, using the results of the valuation survey. Health state classifications amenable to valuation can be developed *de novo* or derived from existing, non-preference-based CSMs. Development of a de novo health state classification requires a procedure that involves interviews with patients in order to identify aspects of HRQoL that are important to them and related to the condition examined, followed by a process of testing and refinement using psychometric methods and focus groups, until the final classification system is developed. The new measure needs to be assessed for its psychometric properties, such as its construct validity and responsiveness. The advantage of such a process is that the new measure can be best suited to the purpose it was constructed for; on the other hand, such a task can be time-consuming and costly. Another drawback of the approach is that the newly developed measure cannot be used in retrospective economic evaluations using existing datasets (Brazier et al., 2007).

Derivation of a health state classification from an existing instrument relies on the selection of a sample of the most representative domains and best performing items within each domain of the original measure and possibly a modification of the item response levels, using a number of psychometric and other statistical methods. Selection of items and response levels is essential so that health states described by the new measure are amenable to valuation;

retaining all items included in the original CSM would likely lead to a large number of health states that consist of multiple statements, which would be impossible to handle in a valuation survey. On the other hand, omitting items from the original measure entails the danger of loss of descriptive information (Brazier et al., 2007). Derivation of a health state classification from an existing CSM may be a more pragmatic approach when an appropriate measure for the condition examined is available. This approach is most useful when the original measure is a validated measure that is widely used in clinical practice and research; in this case, derivation of a PBM from the original measure increases its scope, as it allows not only assessment of clinical effectiveness of interventions and programmes, but also economic evaluation alongside clinical studies. Derivation of a new PBM from an existing CSM is useful when the original measure is more relevant and sensitive to the changes in HRQoL in the study population and more acceptable to patients, clinicians and researchers than a generic measure (Brazier et al., 2007).

The development and use of condition-specific PBMs raises concerns regarding comparability across different conditions and patient populations, which have been (and are still) expressed in an on-going debate (Brazier & Fitzpatrick, 2002; Brazier & Tsuchiya, 2010; Dowie, 2002a & 2002b; Feeny, 2002; Guyatt, 2002). This issue is discussed in more detail in Chapter 8 (section 8.4.1).

1.3.4 Recommendations on the use of preference-based measures in economic evaluation of healthcare interventions

Several regulatory and advisory bodies worldwide (for example in England and Wales, Ireland, Spain, Portugal, Italy, France, Germany, Belgium, the Netherlands, Denmark, Sweden, Norway, Poland, Russia, Canada, US, Brazil, Australia, New Zealand, South Africa, Egypt, China, Taiwan and Thailand) have developed recommendations on the use of PBMs for the estimation of QALYs in cost-utility analyses of healthcare interventions, with several bodies advocating the use of generic PBMs (http://www.ispor.org/PEguidelines/index.asp). The National Institute for Health and Care Excellence (NICE) in England and Wales has explicitly expressed a

preference for the EQ-5D for the estimation of QALYs in cost-utility analyses of healthcare technologies for adults, in order to ensure consistency and comparability across the Institute's appraisal programme. NICE, however, acknowledges that EQ-5D data may be unavailable or inappropriate for the condition or effects of treatment. When EQ-5D data are not available, NICE proposes the use of mapping in order to link available HRQoL measures to EQ-5D utility values; in this case an adequate mapping function needs to be demonstrated and validated. NICE also accepts the use of alternative, standardised and validated PBMs in cases where EQ-5D is unavailable and mapping is not possible or where EQ-5D is inappropriate, but requests that the reason for the use of the alternative measure be fully explained and supported by empirical evidence of its properties. When alternative PBMs are selected for use in cost-utility analysis, the institute recommends that measurement of changes in HRQoL be reported directly from patients, and the respective utility values be based on public preferences, elicited from a representative sample of the UK general population using a choice-based method [i.e. TTO or SG], with 'full health' as upper anchor, so as to retain methodological consistency with the methods adopted at the valuation of EQ-5D (National Institute for Health and Care Excellence, 2013).

1.4 Setting the context for this thesis

1.4.1 The societal burden of mental disorders and current levels of mental health research in the UK

Mental illness is an important cause of disability in the UK. In 2006, there were nearly one million recipients of incapacity benefit due to mental and behavioural disorders, comprising 40% of total incapacity benefit recipients in the country. Over the same year, more than 10 million working days were lost due to stress, depression and anxiety (Oxford Economics, 2007). In England, 8.65 million people were estimated to be suffering from mental health disorders in 2007, incurring costs for their management and associated productivity losses totalling £48.6 billion (McCrone et al., 2008). The respective projected figures for 2026, reflecting an expected increase in the population by 15.1%, were 9.88 million people with mental disorders incurring £60.69 million in 2007 prices (McCrone et al., 2008).

The importance of mental health for the personal and societal well-being has been acknowledged by the Coalition Government, which has set up a strategy for the improvement of mental health outcomes in people of all ages (HM Government & Department of Health, 2011). However, despite the substantial financial and disability burden caused by mental disorders, mental health is an area largely neglected in terms of research in the UK: it has been estimated that, although 15% of disability resulting from disease is attributable to mental illness, only 6% of medical research is currently directed into mental health (Medical Research Council, 2010). This dearth in mental health research is possibly reflected in the extent of clinical and economic appraisal of interventions and programmes for people with mental disorders treated in the NHS in England and Wales: since its establishment in 1999 and up to March 2013, NICE had produced 15 Technology Appraisals and 26 Clinical Guidelines relating to interventions and care pathways for people with mental disorders (available from www.nice.org.uk). The amount of mental health-related guidance is surprisingly limited compared with the total number of Technology Appraisals (277) and Clinical Guidelines (165) published by the Institute over the same period, potentially indicating a lower interest of policy-makers in mental health care compared with other disease areas, or considerable limitations in the quantity and quality of clinical and economic data in the area of mental health, which prevent the development of useful guidance.

1.4.2 Rationale for and objective of this thesis

Mental health is one broad area where the appropriateness of the use of generic PBMs in order to generate QALYs for use in economic evaluation has been questioned (Brazier, 2008 & 2010; Chisholm et al., 1997; Knapp & Mangalore, 2007). This is because generic measures have been primarily designed to capture physical health problems and may miss important aspects of HRQoL of people with mental disorders. For example, EQ-5D focuses on physical health, with only one item addressing mental health problems (depression/anxiety). Due to their limited perceived relevance in this area, generic measures have been found to be less acceptable to patients and clinicians (Crawford et al., 2010; Gilbody et al., 2003), and this result may

explain their limited use in clinical practice and research (Gilbody et al., 2003). The scepticism on the use of generic measures in mental health has led to arguments towards the development of a mental health-specific PBM, which will be relevant to the course of illness of people with mental disorders and sensitive to changes in their HRQoL status (Brazier, 2008; Chisholm et al., 1997; Knapp & Mangalore, 2007).

However, the vast majority of condition-specific measures in mental health are not preference-based (that is, they are not linked to utility values) and are thus not suitable for estimation of QALYs in economic evaluation. Given this gap between the necessity for economic evaluation of mental health interventions in the UK setting and the unavailability of an appropriate PBM that is able to capture important aspects of HRQoL of people with mental disorders, the main objective of this thesis was the development of a preference-based CSM for mental health problems, as a more appropriate and sensitive measure of HRQoL compared with generic PBMs in this area. Considering the plethora of validated CSMs that are available in the area of mental health, the aim was not to develop a new PBM *de novo*, but, instead to derive it from an existing valid, responsive, acceptable and widely used CSM, with the expectation that the new PBM will enable wider assessment of healthcare interventions and programmes for the management of mental health disorders in the form of cost-utility analysis, both prospectively and retrospectively, using historical data.

1.4.3 Key stages and outline of the thesis

In order to achieve the main objective of the thesis, the following key stages of research were undertaken, the methods and results of which are reported in the remaining chapters of this thesis report.

1. A systematic literature review of the use and psychometric performance of generic PBMs in mental health research and practice was carried out first, in order to explore in depth the appropriateness of using generic measures in populations with mental health problems. The methods and results of this piece of research are reported in Chapter 2.

2. A systematic literature review of outcome measures used in mental health research and clinical practice was also undertaken, aiming at identifying appropriate measures with proven validity, sensitivity and acceptability and able to capture a wide range of symptoms and HRQoL aspects that are relevant to people with mental disorders, as candidates for the derivation of a mental health-specific PBM. The most appropriate CSM of those identified in the literature was subsequently determined according to a number of criteria. Chapter 2 provides the details and the results of this review and, ultimately, focuses on the description of the properties and the applications of the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM), which was selected for the derivation of a mental health-specific PBM based on its being a suitable, widely used and acceptable PROM in mental health research and practice, particularly within the British NHS context.

3. A further systematic literature review was conducted to identify and critically appraise methods reported in the literature for the derivation of health state classifications amenable to valuation from existing, longer measures (either generic or condition-specific), aiming to identify appropriate methodologies that might contribute to the derivation of a new health state classification from the CORE-OM. The methods and findings of this review are reported in Chapter 3.

4. A new health state classification was derived from the CORE-OM, following the development and application of new methodology, which was essential because of the nature of the original measure, which was characterised by highly correlated domains. The specific methods developed and applied in order to derive a health state classification from the CORE-OM are described in Chapter 4. Chapter 5 provides detailed results of the process that led to the derivation of a new health state classification named CORE-6D (Clinical Outcomes in Routine Evaluation – 6-dimensional health state classification).

5. A valuation survey using TTO was undertaken to elicit preferences from members of the general population for selected health states described by the CORE-6D. Subsequently, econometric modelling attached utility values to all health states described by CORE-6D, leading to the construction of a preference-based index that can be used for the estimation of QALYs in cost-utility analysis. The methods and results of the valuation survey and the techniques employed for modelling the valuation data are reported in Chapter 6.

6. A series of psychometric tests, statistical analyses and qualitative assessments were conducted to assess the performance of the new PBM in terms of construct and content validity and responsiveness relative to the generic EQ-5D and SF-6D; CORE-6D was also compared with CORE-OM regarding its construct validity and responsiveness, in order to evaluate the degree of loss of information resulting from moving from a 34-item instrument to a 6-item one. The results of these analyses are provided in Chapter 7.

Finally, Chapter 8 discusses the contribution of this thesis to the broader methodology for the derivation of PBMs from existing measures, describes the implications for policy deriving from the development of the new PBM, points to some broader issues on the role of condition-specific PBMs in the wider healthcare resource allocation environment and the role of patients' preferences in the economic assessment of healthcare interventions, and proposes areas for further research.

Chapter 2. Generic and condition-specific outcome measures in mental health. Selecting an appropriate outcome measure for the derivation of a mental health-specific preference-based measure

2.1 Introduction

As discussed in Chapter 1, generic PBMs have been advocated by regulatory and advisory bodies for the estimation of QALYs in cost-utility analysis of healthcare interventions. NICE has expressed a preference on the use of EQ-5D to ensure consistency across its appraisal programme. However, there are concerns that, due to their focus on physical health, EQ-5D and other generic PBMs may not be appropriate for use in the area of mental health, leading to proposals for the use of a mental health-specific PBM in economic evaluations of mental health interventions and programmes (Brazier, 2008 & 2010; Chisholm et al., 1997; Knapp & Mangalore, 2007).

Currently, over 1,400 outcome measures are used in adult mental health research and practice (National Institute for Mental Health in England, 2008). These include measures that have been designed for the identification, assessment or monitoring of a particular mental disorder, and 'generic' mental health measures, which are applicable across a range of mental disorders (National Institute for Mental Health in England, 2008). At the moment, the vast majority of outcome measures in mental health are not preference-based and thus cannot be used for the estimation of QALYs in cost-utility analysis of mental health interventions. A small number of PBMs specific to major mental health conditions have been developed, including the McSad utility measure, a PBM the was developed *de novo* for use in unipolar major depression (Bennett et al., 2000a & 2000b); a vignette-based descriptive system that estimates utility scores for patients with schizophrenia using their scores on the clinician-

rated Positive and Negative Syndrome Scale (PANSS) (Lenert et al., 2004); and two utility measures derived from DEMQOL and DEMQOL-proxy that measure utility of patients with dementia and their carers, respectively (Mulhern et al., 2012b; Rowen et al., 2012). In addition, a small number of studies have reported utility values for mental health state descriptions that have been based on vignettes; such vignettes have been developed, for example, for unipolar depression (Revicki & Wood, 1998; Schaffer et al., 2002), bipolar disorder (Revicki et al., 2005), and schizophrenia (Briggs et al., 2008; Revicki et al., 1996). However, the scope of the above measures and vignette-based utilities is narrow and their use is restricted to the economic assessment of interventions for the specific mental condition for which they were constructed and cannot be expanded to cost-utility analyses in different mental health areas. Ideally, a mental health-specific PBM should be 'generic' in the sense that it should be able to capture a variety of mental health symptoms and aspects of HRQoL across a wide range of mental disorders, thus enhancing comparability across cost-utility analyses conducted in different areas of mental health.

A generic mental health-specific PBM can be developed *de novo*, or derived from an existing CSM for mental health. A *de novo* PBM (i.e. a new health state classification system) can be developed based on in-depth interviews with experts, patients and carers, in order to identify the important aspects of HRQoL under the condition the new measure aims to capture. Such qualitative techniques for the identification and selection of relevant dimensions and items of the new measure ensure its content validity. Subsequently, quantitative psychometric methods can be used to develop and refine the new measure (Brazier et al., 2007). Although this process can lead to the development of a measure with high validity, reliability and responsiveness (concepts that are defined in section 1.2.4 of Chapter 1), developing a *de novo* PROM is a lengthy process. Moreover, a *de novo* PBM must be used in addition to other instruments that measure symptoms and other health-related aspects, and this requires extra time for its completion which may potentially reduce the acceptability of the new measure to both patients and clinicians/researchers.

Nevertheless, as reported above, mental health is an area where a variety of outcome measures is currently being used (National Institute for Mental Health in England, 2008). The breadth of measures that are valid, sensitive and acceptable indicates that it is possible to derive an appropriate PBM from an existing CSM for mental health, provided that the new measure maintains the properties of the original one (Brazier et al., 2007). The new PBM can be applied to datasets containing the original measure without requiring extra time for its completion and with the further advantage that it allows cost-utility analysis based on retrospective datasets containing the original measure (Brazier et al., 2007). A desired property of the existing CSM (and the derived PBM) is to cover a wide range of dimensions, as health state classifications with a narrow coverage of symptoms and/or aspects of HRQoL may create distortions in preferences elicited in a valuation survey; this may occur if, for example, respondents focus on the narrow perspective of the health state description and ignore other aspects of HRQoL or if the new PBM fails to capture side effects and comorbidities. For the same reason, it has been suggested that CSMs selected for the derivation of PBMs describe HRQoL rather than symptoms, as measures describing HRQoL are likely to be broader in coverage of dimensions (Brazier et al., 2012). The policy issues arising from the use of condition-specific PBMs, especially those with narrow coverage of dimensions, are discussed in Chapter 8.

Following the above considerations, the objective of this chapter is three-fold:

a. To systematically review the psychometric properties of generic PBMs in a range of mental disorders, so as to explore whether the concerns expressed regarding the appropriateness of use of generic PBMs in mental health are justified
b. To systematically review outcome measures used in mental health research and practice, aiming at identifying mental health-specific measures with proven validity, sensitivity and acceptability, able to capture a wide range of symptoms and HRQoL aspects that are relevant to people with mental disorders
c. To select one of the CSMs identified from the systematic literature review as the basis for the derivation of a PBM specifically designed for

use across different mental disorders, in particular within the British NHS context.

2.2 Systematic literature reviews: methods and overview of results

One highly specific systematic search of the literature was conducted to identify evidence on the appropriateness of use and psychometric properties of the 3 most widely used generic PBMs (i.e. EQ-5D, SF-6D and HUI-3) in mental health conditions (systematic review 1), and also to review outcome measures used in mental health research and clinical practice aiming at identifying appropriate measures for the derivation of a generic mental health-specific PBM (systematic review 2). The following databases were searched for this purpose:

Via OVID interface

1. EMBASE (1980 to current)
2. MEDLINE
3. PsycInfo
4. Health Management Information Consortium (HMIC)

Via Wiley interface

5. Cochrane Database of Systematic Reviews (CDSR)
6. Cochrane Methodology Register
7. Health Technology Assessment (HTA) database
8. Database of Abstracts of Reviews of Effects (DARE)

The systematic search was initially performed in March 2007. The search was updated in December 2012, after work on the development of the new PBM that was undertaken for this thesis was completed, to provide a more comprehensive picture of the performance of generic PBMs in the area of mental health and of outcome measures used in mental health that would be appropriate to use for the derivation of a new generic mental health-specific PBM. Therefore, some of the retrieved evidence had a more confirmatory rather than exploratory role, since indications on the inappropriateness of generic PBMs for use in mental health conditions and the appropriateness of

the finally selected CSM for the derivation of a mental health-specific PBM were already present prior to the results of the review update.

One common search strategy was developed for the two systematic reviews, which adopted and/or modified search terms included in the search strategies of two reports published by the University of York, both of which explored outcome measurement in mental health research and practice (Gilbody et al., 2003; Jacobs, 2009). Additional search terms for quality of life and generic PBMs were also added, given the extended scope of the search. The search strategy used for the systematic search of the literature is provided in Appendix 2.

The following inclusion/exclusion criteria were applied to select studies identified by the search for further consideration:

- Only literature reviews were included in each of the two reviews; this was decided because a preliminary search had captured a high number of hits (approximately 30,000), and, at the same time, had identified an adequate number of reviews addressing the research questions
- Studies published from 2002 onwards were included, to reflect recent trends in outcome measurement in mental health research and practice
- Only papers published in English language were considered
- Only studies assessing generic PBMs (EQ-5D, SF-6D, HUI-3) or condition-specific outcome measures in adults with mental disorders were included; studies focusing on children and adolescents were excluded from the reviews
- The populations examined in the studies should have a primary diagnosis of a mental disorder in a community, primary, secondary or tertiary setting
- The instruments described and appraised in systematic review 2 should be used for outcome measurement and monitoring of people with mental disorders, including symptoms, functioning and quality of life; measures aiming at case identification were not of interest
- The reviews should focus on the use and properties of generic PBMs or CSMs in the area of mental health, and not on the assessment of the

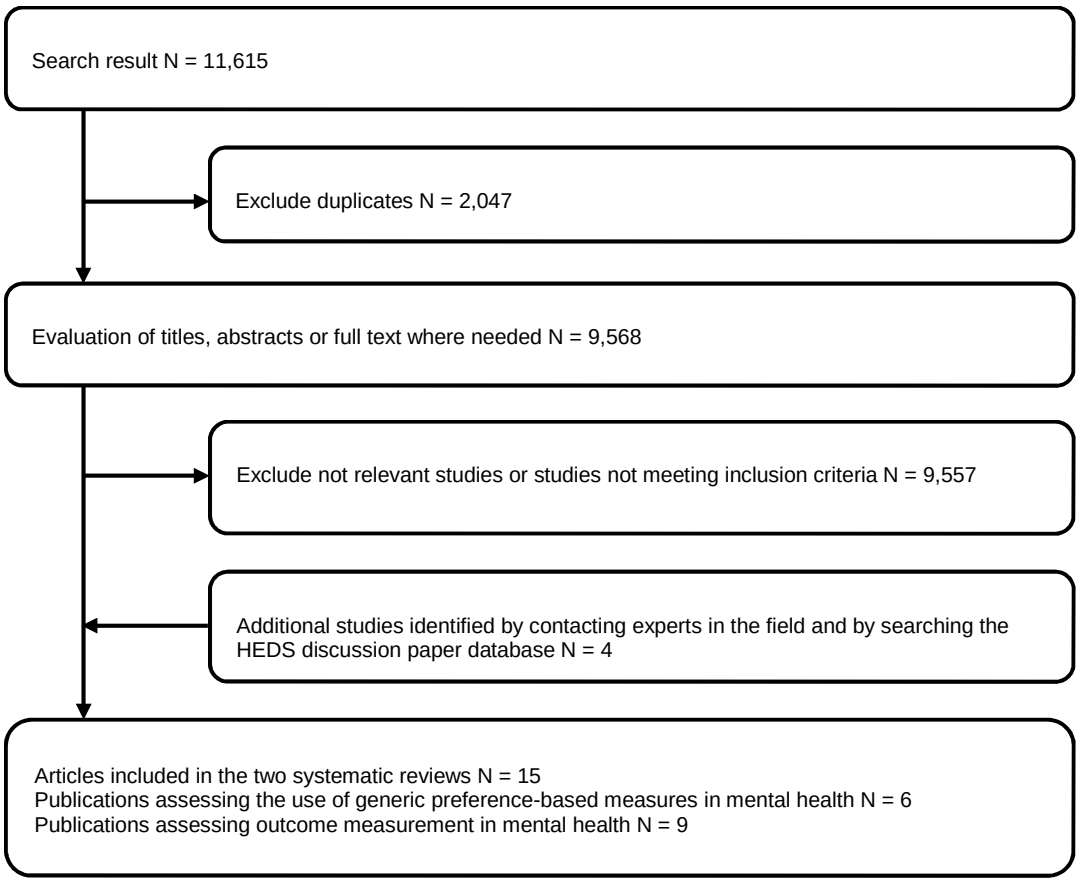
underlying attributes the measures aimed to capture; that is, studies aiming to describe and analyse the symptoms, functioning, HRQoL or overall course of illness of people with mental disorders as reflected in outcome measurement were not included in the reviews

- Conference abstracts and poster presentations were excluded from consideration, as they did not provide sufficient details of methods and results.

In addition to the systematic search of the literature, experts in the field were contacted for potential studies that were on-going or unpublished at the time the final search was performed or for additional reports and studies not identified by the search. The Health Economics and Decision Science (HEDS) discussion paper database of University of Sheffield (http://www.shef.ac.uk/scharr/sections/heds/discussion) was also searched.

The systematic search identified 11,615 references in total. After excluding 2,047 duplicates, 9,568 titles and/or accompanying abstracts were screened for relevance against the set inclusion/exclusion criteria. Full texts of studies potentially meeting inclusion criteria (including those for which eligibility was not clear from the abstract) were obtained. Moreover, 4 additional studies meeting inclusion criteria were identified by contacting experts in the field and by searching the HEDS discussion paper database. After excluding studies clearly not relevant to the topic and studies not meeting inclusion criteria, 15 publications were included in the review, consisting of 6 publications reviewing the performance and properties of generic PBMs in mental health and 9 publications reviewing outcome measurement in mental health research and practice. A flow diagram showing the systematic process for selecting papers for the review is provided in Figure 4.

Figure 4. Flow diagram of the selection of publications in the systematic review of generic preference-based measures and outcome measures used in mental health research and practice



2.3 The appropriateness of using generic preference-based measures in mental health – results of systematic review 1

2.3.1 Overview of search results

The systematic search of the literature identified 6 review publications reporting on the appropriateness and psychometric properties of generic PBMs in mental health populations. Five of these studies examined the psychometric properties of EQ-5D and SF-6D, and one focused specifically on EQ-5D. No review assessing the appropriateness and psychometric properties of HUI-3 in the area of mental health was identified by the search.

One of the seven studies identified (Brazier et al., 2014) was a large study with many components that were relevant to this review, the majority of which have been published separately in other publications identified by the search. The study by Brazier and colleagues (2014) included a systematic review aiming to assess the precision, construct validity and responsiveness of generic measures (EQ-5D, SF-36, SF-12 and SF-6D) in five mental health conditions, that is, depression and anxiety (with detailed results being reported by Peasgood and colleagues, 2012); schizophrenia (detailed results of which have been reported by Papaioannou and colleagues, 2011); personality disorders (detailed results reported by Papaioannou and colleagues, 2013); and bipolar disorder. Construct validity was assessed in terms of known groups validity and convergent validity. Responsiveness to change over time was measured by the ES, SRM, and the correlation of the generic measure's change score with change scores of other measures that were considered relevant in measuring symptoms and/or HRQoL aspects associated with the underlying mental condition.

In addition to the quantitative assessment of the psychometric properties of generic measures, Brazier and colleagues (2014) examined the content validity of generic measures in the area of mental health. For this purpose the authors carried out a systematic review of qualitative studies reporting the views of people with mental health problems on mental symptoms and related aspects of HRQoL. Subsequently, the authors undertook framework analysis

to identify common patterns of themes within and across different qualitative studies, so as to establish themes of HRQoL that are important to people with mental disorders. The items of generic measures were subsequently compared, in a qualitative manner, against these themes. The results of this review have also been made available by Connell and colleagues (2012).

Further to this large study by Brazier and colleagues (2014) and its separately published components, the systematic search conducted for this thesis identified one review examining the psychometric properties of EQ-5D (feasibility, precision, construct validity and reliability) in dementia, regarding both patient and proxy (carers' or clinicians') ratings (Hounsome et al., 2011). Feasibility of EQ-5D was assessed by the rates and time of completion. Precision was examined by the level of ceiling and floor effects. Construct validity was assessed in terms of known groups validity and convergent validity. Reliability was assessed by test-retest reliability. Agreement (correlation) between patient and proxy ratings as well as between ratings of different proxies was also measured.

It must be noted that the objective of the review of reviews conducted for this thesis was to evaluate the appropriateness and psychometric properties of generic PBMs. Thus, only data relating to the properties of EQ-5D and SF-6D (regarding the health state classifications and the utility indices) were extracted from the existing reviews. Reported data on EQ-VAS (a VAS administered alongside EQ-5D for recording an individual's rating of their current state of HRQoL), SF-36 or SF-12 were not considered. Table 1 provides an overview of the review studies considered in this review of reviews, including the mental disorders and the generic PBMs that were examined in each review, the psychometric properties assessed and a summary of each study's findings.

Table 1. Studies included in the systematic review of reviews examining the properties of generic preference-based measures in mental health populations – overview of study methods and results

Review study reference	Mental health area	Generic PBM	Psychometric properties assessed and overview of findings	
Brazier et al., 2014; Peasgood et al., 2012	Depression and anxiety	EQ-5D [k=21] SF-6D [k=8]	Known groups validity Convergent validity Responsiveness	good - both measures strong in depression; moderate in anxiety (moderate correlation to depression measures, small correlation to anxiety measures) - both measures good - both measures
Brazier et al., 2014; Papaioannou et al., 2011	Schizophrenia	EQ-5D [k=9] SF-6D [k=1]	Known groups validity Convergent validity Responsiveness	good - EQ-5D mixed evidence - EQ-5D; moderate - SF-6D mixed evidence - EQ-5D; poor - SF-6D
Brazier et al., 2014; Papaioannou et al., 2013	Personality disorders	EQ-5D [k=6]	Known groups validity Convergent validity Responsiveness	moderate moderate moderate to high
Brazier et al., 2014	Bipolar disorder	EQ-5D [k=4]	Known groups validity Convergent validity Responsiveness	mixed evidence mixed evidence no evidence
Hounsome et al., 2011	Dementia	EQ-5D [k=17; 14 on self-ratings, 15 on proxy ratings]	Feasibility Precision Known groups validity Convergent validity Test-retest reliability Agreement	good in mild dementia, low to moderate in more severe dementia - self-ratings; good - proxy rating ceiling effects - self-rating poor - self-rating; good - proxy rating poor - self-rating; good - proxy rating good in mild and moderate dementia - self-rating; overall good - proxy rating poor between self-/proxy ratings and between different proxies
Brazier et al., 2014; Connell et al., 2012	General	EQ-5D [NA] SF-6D [NA]	Content validity	lack of coverage of various relevant aspects of HRQoL

*k = number of primary studies assessing each generic PBM in each review

2.3.2 Psychometric properties of EQ-5D and SF-6D in people with mental disorders

Depression and anxiety

The systematic literature review on the properties of generic measures conducted by Brazier and colleagues (2014) identified 26 primary studies on patients with a primary diagnosis of depression and/or anxiety. Of these, 21 studies reported data on EQ-5D and 8 studies included SF-6D data.

According to the review findings, which have been reported in detail by Peasgood and colleagues (2012), EQ-5D was able to identify a reduction in HRQoL in people with depression and anxiety, and could detect significant changes in HRQoL across different symptom severity levels. The loss of HRQoL was more evident in the domains of depression and anxiety, pain and discomfort, and usual activities, and less so in mobility and self-care. In people with depression EQ-5D correlated strongly with clinician-rated measures of depression severity and moderately with measures of functioning, patient-reported severity and patient-reported quality of life. In people with anxiety EQ-5D had moderate correlations to depression measures (such as the Back Depression Index -BDI-) but only small to moderate correlations to anxiety measures. EQ-5D was very responsive in both depression and anxiety, with a similar degree of responsiveness to symptom, functioning and quality of life measures.

SF-6D was also able to detect loss of HRQoL in people with major depression and anxiety disorders and to distinguish across different symptom severity groups. For people with depression, SF-6D captured a considerable loss in HRQoL in the domains of mental health, vitality, role limitations, social functioning and bodily pain and a lower HRQoL loss in physical functioning. SF-6D was found to correlate well with the Patient Health Questionnaire – 9 items (PHQ-9), a measure of depression, but less so with anxiety scales. SF-6D was found to be responsive in major depression and in a mixed population of people with depression and/or anxiety.

Comparison between EQ-5D and SF-6D in studies that used both measures revealed that SF-6D was better at capturing mild depression and anxiety, whereas EQ-5D was better in identifying more severe symptom levels.

Conclusively, EQ-5D and SF-6D showed good known groups validity in people with depression and anxiety. Both measures demonstrated good convergent validity in people with depression; in people with anxiety, EQ-5D and SF-6D were able to capture changes in HRQoL relating to depressive symptoms or comorbid depression but were less effective in tapping anxiety symptoms. Both measures showed high responsiveness.

Schizophrenia

Brazier and colleagues (2014) also conducted a systematic literature review to assess the construct validity and responsiveness of EQ-5D, SF-36, SF-12 and SF-6D in people with schizophrenia, the results of which have been reported in detail by Papaioannou and colleagues (2011). The review included 33 primary studies; of these, 9 assessed the psychometric properties of the EQ-5D and only one study assessed the validity of SF-6D.

EQ-5D had good known groups validity as it was able to distinguish between patients with ‘severe’ or ‘less severe’ symptomatology, defined by various CSMs including the PANSS, the Hamilton Depression Rating Scale (HAM-D) and the Global Assessment of Functioning (GAF). Regarding convergent validity, EQ-5D appeared to correlate moderately to strongly with the Health of the Nation Outcome Scales (HoNOS), the Symptom Checklist-90-Revised (SCL-90-R), the Brief Psychiatric Rating Scale (BPRS), the Schizophrenia Quality of Life questionnaire (S-QoL), and the Clinical Global Impression Severity of Illness Scale (CGI-S), and weakly to moderately with the Global Assessment of Relational Functioning (GARF). There was also limited evidence suggesting moderate to strong associations between EQ-5D and depression or anxiety symptom measures; however, there was mixed evidence for the measure’s correlation with PANSS, GAF and the Social and Occupational Functioning Assessment Scale (SOFAS), with some evidence suggesting moderate to strong correlation with these measures and other

evidence indicating weak to non-existent association. EQ-5D was not correlated to the Quality of Life Scale (QLS), a schizophrenia-specific quality of life measure. Data on responsiveness were also mixed: EQ-5D was responsive to change over time, but at the same time it did not respond to changes in most symptom or functioning measures, apart from significant correlations with a few measures, including the PANSS positive subscale. Moreover, EQ-5D changes were found not to correlate with BPRS changes, unless the latter were greater than 25%.

The limited evidence for SF-6D indicated that the measure correlated moderately with BPRS; however, SF-6D change scores correlated only weakly with BPRS changes, and only when the latter were greater than 25%.

Overall, evidence suggests a good known groups validity but rather inadequate convergent validity of generic PBMs in schizophrenia. The responsiveness of EQ-5D and SF-6D is not satisfactory either. Generic PBMs are probably unable to adequately capture changes in HRQoL of people with schizophrenia.

Personality disorders

Brazier and colleagues (2014) conducted a systematic review of the construct validity and responsiveness of generic measures (EQ-5D, SF-36, SF-12 and SF-6D) in people with personality disorders, the detailed results of which have been reported by Papaioannou and colleagues (2013). The review included 10 studies, 6 of which included EQ-5D data; none of the studies assessed SF-6D.

EQ-5D was able to capture HRQoL loss in people with a borderline, narcissistic, obsessive-compulsive, depressive, negativistic, or mixed personality disorder; however, it was the number of personality disorders rather than their type per se, that made a significant effect on the EQ-5D score. When controlling for the number of the disorders, only depressive personality disorder appeared to have a significant effect on EQ-5D. EQ-5D correlated moderately with the Global Severity Index (GSI) and showed moderate to high responsiveness to change over time; EQ-5D change scores

correlated well with change scores on the Borderline Personality Disorder Severity Index-IV (BPDSI-IV), a measure of the severity of borderline personality disorder.

The limited available evidence overall indicates that EQ-5D is a valid and responsive measure in people with personality disorders.

Bipolar disorder

The systematic review of the psychometric properties of generic measures (EQ-5D, SF-36, SF-12 and SF-6D) conducted by Brazier and colleagues (2014) included 22 studies; of these, 4 studies contained data on EQ-5D and no study reported data on SF-6D. Evidence was mixed with some findings supporting the known groups and convergent validity of EQ-5D and some findings questioning it. There was no available evidence on the responsiveness of EQ-5D in people with bipolar disorder.

Dementia

Hounsome and colleagues (2011) carried out a systematic review to assess the acceptability, precision, validity and reliability of EQ-5D in people with dementia and their carers. The review included 17 papers that focused on patients with dementia, 14 of which used the EQ-5D for patient self-assessment, and 15 for proxy assessment by family carers, institutional carers or healthcare professionals; in 12 studies, EQ-5D was used for both self- and proxy assessments.

According to the results of the review, feasibility of EQ-5D was high in people with mild dementia, as demonstrated by high completion rates, but low to moderate in people with moderate or severe dementia. The mean completion time of self-reported EQ-5D ranged from 4 minutes to more than an hour. In patients with mild and moderate dementia, the self-reported EQ-5D showed good to average test-retest reliability, but lower than carers' proxy ratings. A good proportion of patients (more than one-third) rated their HRQoL at the highest level for several or all EQ-5D dimensions, indicating a potential ceiling effect for EQ-5D. Patient-reported EQ-5D was not able to distinguish across

different severity levels of dementia. Of the 5 EQ-5D dimensions, only depression/anxiety had a positive correlation with the Mini-Mental State Examination (MMSE), a scale measuring cognitive impairment. EQ-5D correlated better with measures of depression and anxiety, rather than with physical activity and cognitive function.

There was evidence for good feasibility, reliability and construct (known groups and convergent) validity of EQ-5D proxy ratings. Such ratings correlated well with other HRQoL measures such as the Quality of Well Being scale (QWB) and HUI, and also with patients' cognitive function. However, different proxies provided different ratings: EQ-5D ratings provided by clinicians had higher construct validity in the mobility and self-care dimensions; ratings provided by carers had higher construct validity in usual activities and depression/anxiety dimensions. The level of agreement between carers and physicians was poor, especially for the dimensions of pain and depression/anxiety. On the other hand, patient self-assessment provided significantly higher EQ-5D scores than proxy assessments, and this discrepancy was not attributable to cognitive impairment alone.

Based on the results of their review, the authors concluded that, despite the feasibility and reliability of EQ-5D (both self- and proxy rated), there were problems with the validity of self-rated data, demonstrated by lack of association between patient and proxy ratings; moreover, there were important discrepancies in ratings among different proxies, making interpretation of EQ-5D scores in the area of dementia problematic.

2.3.3 Qualitative evidence – content validity of generic preference-based measures in people with mental disorders

The study by Brazier and colleagues (2014) also included a systematic review of qualitative research undertaken on people with mental health problems, aiming to identify the domains of HRQoL that are important to this population. The methods and the results of this review have been made available in Connell and colleagues (2012). The methods of this review are described in more detail here because this comprised a high quality synthesis of qualitative

research, and its findings formed the basis for the assessment of the content validity not only of generic PBMs in the area of mental health, but also of the new mental health-specific PBM that was developed for this thesis. The systematic review included 13 studies that gathered qualitative evidence through interviews and focus groups; in these primary studies adults with mental health problems were explicitly asked what they considered to be important to their quality of life or how their quality of life had been affected by their mental health problems. Findings across the 13 studies was synthesised using a framework approach, which comprises a highly structured method for organising and analysing data that allows the expansion and refinement of an *a-priori* framework to incorporate new themes that emerge from the data (Connell et al., 2012). Framework analysis allowed the identification of common and variable patterns of themes within and across different studies and consisted of five stages: familiarisation with the studies; identification of initial themes for a thematic framework; data organisation; examination of each initial theme and identification and documentation of further sub-themes within the framework chart; and mapping across the sub-themes in order to make connections between them, and assist in the development of the final themes (Connell et al., 2012).

This work identified six domains that were important to people with mental disorders: well-being and ill-being; control, autonomy and choice; self-perception; social well-being, belonging and relationships; activity and functioning; hope and hopelessness. Each domain includes positive and negative aspects. A seventh theme, that of physical health, was identified by direct interviews with users of mental health services.

Well-being was defined by high levels of pleasant emotions and moods and low levels of unpleasant ones. Ill-being was defined by general feelings of distress, experience of psychotic or manic symptoms (including hallucinations and delusions, reality disorientation, mania, discomfort, weirdness and irritability), depressed mood, fear and anxiety, and problems with energy and motivation.

Control referred to the availability of external resources which enabled choice and control, including medication and treatment, support, information and finances. An important issue for people with mental disorders was the relief and management of most distressing symptoms of their condition, achieved mainly through medication. People with mental illness found important being informed and having an insight about the condition and what to expect from it in the future, as well as being able to develop strategies to effectively manage their illness. A balance between support and independence was deemed important for the autonomy of people with mental disorders. Choice was raised in the context of availability of financial resources and employment opportunities.

Self-perception was identified as an important aspect of HRQoL, in terms of self-efficacy (a belief and confidence in own abilities as opposed to feelings of uselessness, failure and helplessness), self-identity (a good perception of own self), and self-esteem (having a sense of self-worth and self-respect), with all concepts being linked to self-acceptance (as opposed to self-stigmatisation and not feeling normal).

Social well-being and belonging comprised the need for integration within the social environment and the experience of feeling valued, needed and accepted. Relationships, including family and friends as well as social relationships in the community, played a central role in this social integration and acceptance. At the opposite end, stigmatisation and the perception of negative reactions from family, friends and the community was highlighted as placing a considerable burden to people with mental disorders.

Activity and functioning were identified as an important aspect of HRQoL and included both employment and leisure activity. Participating in an activity was deemed to help achieving a sense of self, interacting with others and giving a sense of belonging and participating in the external world. Activity also improved mood, provided a distraction from problems, increased self-esteem, provided routine and structure, and enabled people to take control of their lives.

Hope was defined as having dreams and goals, having meaning and purpose, and moving forward in life. On the opposite end, past losses, including the loss of life roles in general, and the loss of work, relationships, skills, time, finances, and, ultimately, the loss of self-identity, led to a pervasive feeling of distress and hopelessness, characterised by a view that life would never change for the better, which placed a burden upon the HRQoL of people with mental disorders..

The review revealed the complexity of the factors that determine the HRQoL in people with mental health problems and the difficulty in separating mental symptoms from other aspects of HRQoL in this population. The authors acknowledged as a limitation of the review the fact that the included studies focused on the HRQoL of people with severe mental illness, particularly schizophrenia. However, further interviews with people with a wide range of mental health problems, including milder mental illness, confirmed these findings.

The evidence from this review was used to determine the content validity of generic measures EQ-5D, SF-36 and SF-6D and of a new generic measure of capability in adults, the ICECAP-A, in the study by Brazier and colleagues (2014). The content validity of each measure was judged by the extent to which its items of each measure represent the areas that have the greatest impact on the HRQoL of people with mental health problems.

The assessment of the content validity of generic measures in the area of mental health against the 7 domains of HRQoL that were identified as important to people with mental disorders concluded that EQ-5D covered little of the content of these 7 domains due to its focus on physical health. Of the 7 domains EQ-5D captures well the one on physical health (by items on mobility, self-care, usual activities and pain). Activity and functioning is roughly covered by the EQ-5D item on usual activities. Social well-being, belonging and relationships is only partially captured by the EQ-5D item on usual activities. Subjective ill-being (but not well-being) is broadly reflected in the EQ-5D item

on depression and anxiety. EQ-5D is not able to represent the remaining 3 domains of HRQoL that are important to people with mental disorders, that is, control, autonomy and choice; self-perception; and hope and hopelessness.

Compared with EQ-5D, SF-6D covers a wider range of the 7 HRQoL domains that are relevant to mental health, as it is more balanced in tapping both physical and mental health aspects. SF-6D captures the domain of physical health through its questions on physical functioning, vitality, and bodily pain and its sub-question on role limitations due to physical problems. Activity and functioning is covered by SF-6D items on physical functioning and role limitations. Social well-being, belonging and relationships is broadly captured by the social functioning item of SF-6D. Subjective well-being and ill-being is covered by the mental health and the vitality SF-6D items and, partly, by the sub-question on role limitations due to emotional problems. Similar to EQ-5D, SF-6D is unable to capture the remaining 3 domains of HRQoL that are important to people with mental disorders, that is, control, autonomy and choice; self-perception; and hope and hopelessness.

Although Brazier and colleagues (2014) did not evaluate the content validity of HUI-3 in mental health populations, it was possible to do so for this thesis using the themes and the approach described in that report. HUI-3 covers physical health aspects by its items on vision, hearing, pain, ambulation and dexterity. The domain on activity and functioning is only indirectly covered by the HUI-3 items on vision, hearing, speech, ambulation, dexterity, cognition; however, these items cover mainly physical and not mental aspects of activity and functioning. HUI-3 items on vision, hearing and speech may affect social well-being, belonging and relationships, but again they capture only physical aspects that may interact with this domain. Similarly, most HUI-3 items, including hearing, vision, speech, ambulation, dexterity and cognition affect the domain control, autonomy and choice, but basically capture physical factors affecting control and autonomy. The domain of subjective well-being and ill-being is partially captured by the HUI-3 item on emotion. HUI-3 is unable to capture self-perception as well as hope and hopelessness.

The conclusion from the assessment of the content validity of generic PBMs is that these measures are not able to fully capture the HRQoL aspects that are most important to people with mental disorders.

2.3.4 Conclusion on the appropriateness of using of generic preference-based measures in mental health

The review of the psychometric properties of generic PBMs suggest that EQ-5D and SF-6D perform satisfactorily in depression, but less so in anxiety and personality disorders (for the latter, only evidence on the EQ-5D was available). Results were mixed in schizophrenia and bipolar disorder. Results suggest that generic PBMs may be picking depressive symptoms (or comorbid depression) rather than core symptoms associated with a range of conditions, including anxiety and schizophrenia. In dementia, self-reported EQ-5D performs satisfactorily on some psychometric tests (feasibility, reliability) but its validity is questionable. The validity of proxy EQ-5D ratings seems to be higher, but there is poor agreement between patient- and proxy ratings as well as ratings across different proxies. The review of qualitative evidence on important aspects of HRQoL for people with mental health problems leads to the conclusion that generic PBMs fail to address the complexity of quality of life measurement and the broad range of domains that are important to people with mental health problems.

The systematic search for reviews did not identify any evidence on the psychometric properties of HUI-3 in people with mental disorders. Regarding the other two generic PBMs, EQ-5D and, in particular, SF-6D, available evidence varied from limited to non-existent in some areas, and therefore safe conclusions could not be always drawn. Moreover, a number of mental disorders such as Obsessive Compulsive Disorder (OCD), panic disorder, generalised and specific phobias were not covered in the review, as no relevant evidence was identified. Another limitation of the findings is that validity and responsiveness were mostly assessed by correlations of the generic PBMs with other measures that were considered relevant in measuring symptoms and/or HRQoL aspects associated with the underlying mental condition. However, not all of these measures that were used as ‘gold standards’ have proven validity and responsiveness in populations with mental

problems, and therefore results need to be interpreted with caution. Nevertheless, results indicate that generic PBMs may not be appropriate to use for the assessment of HRQoL in people with mental health problems.

In addition to the findings of this review of reviews, one systematic review on outcome measurement in mental health reported that only a negligible portion of Randomised Controlled Trials (RCTs) conducted in mental health research (approximately 1%) use generic measures, although it needs to be acknowledged that this review was published a decade ago (Gilbody et al., 2003). The authors attributed their finding to the focus of generic measures on physical functioning with questions that are irrelevant to people with mental health problems, whilst aspects of social and role functioning that are important to them are ignored, resulting in generic measures being probably insensitive to underlying change in health status and also unacceptable to respondents. Besides, use of generic instruments in addition to CSMs creates extra burden to clinicians and patients alike. Based on their findings, the authors concluded that "*there is no robust research evidence to support the value* [of generic measures] *as routine measures of outcome in psychiatric settings*".

Thus, as Brazier and colleagues (2014) argue, "*overall findings suggest that there seems to be a case for developing a new preference-based measure specific to mental health*". The authors acknowledge that "*it may not be possible to cover all dimensions of physical and mental health with the same level of coverage in one measure, but the new measure would need to incorporate the impact of both physical and mental health problems*".

Therefore, the rest of this chapter aims to identify an appropriate mental health-specific measure with proven validity, sensitivity and acceptability and a wide coverage of mental health symptoms and aspects of HRQoL to be used as the basis for the derivation of a PBM specifically designed for use in mental health.

2.4 Outcome measurement in mental health research and clinical practice – results of systematic review 2

2.4.1 Overview of search results

The systematic search of the literature identified 9 studies that reviewed available outcome measures in the area of adult mental health. Of these, 5 studies examined outcome measures used in any area of mental health, either generic mental health measures, that could be used across different mental disorders, or applicable to specific disorders (Flynn, 2002; Gilbody et al., 2003; Hampson et al., 2011; Jacobs, 2009; National Institute for Mental Health in England, 2008), one study reviewed outcome measures used in major depression (Flynn, 2004) and 3 studies reviewed outcome measures used in schizophrenia (Burns & Patrick, 2007; Flynn, 2003; McCabe et al., 2007). It must be noted that, given the objective of this review, which was to ultimately identify an appropriate CSM as the basis for the derivation of a generic mental health PBM, its focus was on generic mental health measures, that is, on measures that are applicable across a wide range of mental disorders. Nevertheless, studies reviewing outcome measures used in specific mental disorders such as depression and schizophrenia were still considered in the review, because it was possible that they examined mental health measures that are applicable to a wider range of mental disorders and not only to those particular disorders examined in these studies. Furthermore, the review gave higher emphasis to UK-based studies, as one of the desired properties of the CSM was to be widely used in the UK mental health research and practice.

Table 2 provides an overview of the studies included in the review, their potential focus on specific mental disorders, the aims of each study that are relevant to this review and the methodology adopted in each of them. In the text reporting the detailed findings of this review of reviews that follows, all outcome measures that are shown in bold characters are those that

a. were identified in each review as most commonly used or most appropriate for use in mental health populations (depending on the review's objective)

b. can be used across a range of mental disorders (and may thus be considered as generic mental health measures) and
c. have wide enough scope in terms of capturing various aspects of HRQoL.

Table 2. Studies included in the systematic review of outcome measurement in mental health

Study reference	Focus on specific mental disorder?	Summary of study aims relevant to this review and methodology
Burns & Patrick, 2007	Yes – Schizophrenia	**Aim:** to identify outcome measures used most frequently to assess social functioning in schizophrenia and to assess their psychometric properties **Methodology:** systematic literature review
Flynn, 2002	No	**Aim:** to identify suitable generic measures of mental health status, psychiatric symptoms and functioning for use by the US Veterans Health Administration (VHA) mental health services **Methodology:** systematic literature review and application of psychometric & practicality criteria
Flynn, 2003	Yes – Schizophrenia	**Aim:** to identify suitable outcome measures appropriate for schizophrenia for use by the US Veterans Health Administration (VHA) mental health services **Methodology:** review of existing compendia and reviews and application of psychometric & practicality criteria
Flynn, 2004	Yes - Major depression	**Aim:** to identify suitable outcome measures appropriate for depression for use by the US Veterans Health Administration (VHA) mental health services **Methodology:** review of existing compendia and reviews and application of psychometric & practicality criteria
Gilbody et al., 2003	No	**Aim:** to explore the most widely used outcome measures in psychiatric research and UK psychiatric routine practice, with particular reference to patient-reported outcomes **Methodology:** systematic literature review and survey of UK consultant psychiatrists
Hampson et al., 2011	No	**Aim:** to provide guidance on the use of outcome measures in mental health based on what is of clinical value to patients and clinicians and what is feasible in practice **Methodology:** literature review of government and other national reports
Jacobs, 2009	No	**Aim:** to identify the most commonly used outcome measures in the UK mental health services that would be suitable to convert into a preference-based measure, based on a number of set criteria **Methodology:** systematic literature review and interviews with policymakers, academics and NHS staff involved in outcome measurement in adult mental health services
McCabe et al., 2007	Yes - Schizophrenia	**Aim:** to identify patient-reported outcome measures in schizophrenia and to assess their psychometric properties **Methodology:** non-systematic literature review
National Institute for Mental Health in England, 2008	No	**Aim:** to provide a comprehensive list of mental health outcome measures, their use, properties, advantages and disadvantages, in order to support health professionals and inform service users and carers **Methodology:** literature review, consultation with stakeholders and development of a scoring system assessing each measure's quality

2.4.2 Studies evaluating outcome measurement in mental health with a focus on UK research and clinical practice

Gilbody and colleagues, 2003

The aim of this report was to explore outcome measurement in psychiatric research and practice, with particular reference to PROMs. For this purpose the authors carried out a systematic review to identify the most commonly used outcome measures in randomised clinical trials in psychiatry conducted between the years 1956-2000. In addition, in order to identify the most commonly used outcome measures in UK routine practice they conducted a survey of UK consultant psychiatrists. Based on the results of their review, the authors classified outcome measures used in RCTs into 6 categories:

a. Psychopathological rating scales, measuring predominantly symptoms; these were the most commonly used outcome measures in psychiatric research
b. Global outcome measures, which measure the overall (global) severity of the disorder or its impact on overall functioning; such measures were used in less than half of RCTs in psychiatry
c. Generic PROMs, e.g. EQ-5D and SF-36; only 1% of the trials had used such measures
d. Disease-specific PROMs, which examine various domains of HRQoL and are relevant to specific patient groups or disease areas; such measures were used approximately in 2.5% and 16% of RCTs evaluating drugs and psychosocial interventions, respectively
e. Domain-specific PROMs, examining a specific domain associated with HRQoL; around 6% and 30% of RCTs evaluating drugs and psychosocial interventions, respectively, used this type of outcomes
f. Other outcomes: these may include relapse, mortality, service use, etc.

Table 3 presents the outcome measures most commonly used in psychiatric research and UK routine psychiatric practice according to Gilbody and colleagues (2003).

Table 3. Outcome measures most widely used in psychiatric research and UK clinical practice (Gilbody et al., 2003)

A. Psychiatric research (systematic reviews of 490 RCTs conducted between 1956 – 2000)		
Type of outcome measure		**Most widely used outcome measures**
Psychopathological rating scales		Schizophrenia and related disorders: BPRS, PANSS Depression and related disorders: HDRS
Global outcome measures		GAF, GAS
Generic patient-reported outcome measures		SF-36
Disease-specific patient-reported outcome measures		QLS, QOLI, OQLQ
Domain-specific patient-rated outcome measures	Social functioning	SAS, KAS, SFS, REHAB scale
	Role functioning	ADL, KPS
	Perceptions of wellbeing	RSES
	Satisfaction	CSQ
	Physical Health	PDI

B. Psychiatric routine practice (survey of 340 UK consultant psychiatrists)			
Type of mental disorder	**Purpose of measurement**		
	Case identification & severity assessment	Social functioning, quality of life, assessment of patient needs	Assessment of clinical change over time & therapeutic response
Depressive and anxiety disorders	BDI, HADS, HDRS	HoNOS, SAS, SFS	BDI, HADS, HDRS, HoNOS
Cognitive impairment	MMSE	HoNOS	MMSE, HoNOS
Psychotic illnesses	PANSS, HoNOS, BPRS	PANSS, BPRS, HoNOS	PANSS, BPRS, HoNOS
drugs and alcohol problems	CAGE questionnaire	HoNOS	HoNOS

ADL: Katz Index of Independence in Activities of Daily Living; BDI: Beck Depression Inventory; BPRS: Brief Psychiatric Rating Scale; CSQ: Client Satisfaction Questionnaire; GAF: Global Assessment of Functioning; GAS: Global Assessment Scale; HADS: Hospital Anxiety and Depression Scale; HDRS: Hamilton Depression Rating Scale; HoNOS: Health of the Nation Outcomes Scales; KAS: Katz Adjustment Scale; KPS: Karnofsky Performance Scale; MMSE: Mini Mental State Examination; OQLQ: Oregon Quality of Life Questionnaire; PANSS: Positive And Negative Syndrome Scale; PDI: Pain and Disability Index; QLS: Heinrich's Quality of Life Scale; QOLI: Lehman Quality of Life Interview; RSES: Rosenberg's Self Esteem Scale; SAS: Social Adjustment Scale; SFS: Social Functioning Scale

Looking at Table 3, it appears that there are a number of outcomes that can be considered generic mental health measures, in the sense that they can be used across different mental disorders. These include **BPRS**, **HoNOS**, **GAF** and its precursor Global Assessment Scale (**GAS**). Domain-specific measures, although applicable across different disorders, were not deemed comprehensive enough, as they focus on one domain of HRQoL, and thus were not considered as candidates for derivation of a generic mental health PBM. It must be noted that the measures listed as most commonly used in psychiatric research and UK routine practice by Gilbody and colleagues (2003) were not subject to any assessment in the report.

Mental Health Outcomes Compendium

The Mental Health Outcomes Compendium (National Institute for Mental Health in England, 2008) comprises a comprehensive, though not exhaustive, collection of outcome measures that can be used across adult mental health services, aiming to support clinicians and inform service users and carers. The report provides information on a range of available measures in mental health practice, their use, properties, advantages and disadvantages, thus allowing stakeholders to make an informed choice. The list of measures was compiled based on a literature review that identified the most popular and evidenced outcome measures and further consultation with stakeholders to identify additional instruments that were of clinical value or were recommended by service users; subsequently a scoring system was developed to summarise the quality of the included measures in terms of their psychometric properties (such as validity, reliability and responsiveness), stakeholders' priorities (including clinical utility, appropriateness and acceptability), the existing evidence base, and the measure's availability (determined by practicality, training requirements, copyright issues or permissions for their use and associated costs).

The compendium included 188 measures in total. Based on their quality scores and stakeholders' recommendations, 69 measures were shortlisted, for which more detailed information on their properties and use was provided in the report. The shortlisted measures were organised in 18 distinct diagnostic / therapeutic areas, which can be further grouped in 5 broad 'themes'

representing aspects of health and health care, and these are presented in Table 4.

The authors of the report acknowledged that some measures belonged to more than one diagnostic / therapeutic area. Furthermore, a number of measures fitted also in areas beyond the 18 areas reported; for example, the authors expressed the view that some of the shortlisted instruments, such as the Camberwell Assessment of Need Short Appraisal Schedule (CANSAS), the Clinical Outcomes in Routine Evaluation (CORE) measurement tools, the Functional Assessment of the Care Environment (FACE), the HoNOS and the Threshold Assessment Grid (TAG) were essentially global severity measures that were not restricted to exclusively measuring 'outcome of psychological therapies' (CORE measurement tools) or 'health care and needs assessment' (CANSAS, FACE, HoNOS, TAG) and could in practice be used across different diagnostic / therapeutic areas.

Looking at the identified areas in Table 4, it appears that CSMs belonging in the broader themes of 'social functioning and overall well-being' and 'services' are more likely to comprise appropriate candidate measures for the derivation of a generic mental health PBM, in particular those comprising global severity measures, such as CANSAS, CORE measurement tools, FACE, HoNOS and TAG. It must be noted that, of the global severity measures, the **CORE measurement tools**, **HoNOS** and **TAG** were shortlisted in the Compendium based on both a high quality score and stakeholders' recommendations.

Table 4. Outcome measures included in the Mental Health Outcomes Compendium, shortlisted according to their quality score and/or stakeholders' recommendations (National Institute for Mental Health in England, 2008)

Aspects of health and health care	Diagnostic / therapeutic areas	Outcome measures
Mental disorder areas and/or symptoms	Addictions	Addiction Severity Scale/Index; AUDIT; The Severity Dependence Scale; Maudsley Addiction Profile
	Anxiety and depressive disorders	Amritsar Depression Inventory; BDI; Beck Hopelessness Scale; Centre for Epidemiological Studies Depression Scale; EPDS; Fear Questionnaire; GAD-7; Geriatric Depression Scale; GHQ-12; HADS; HAI; Liebowitz social anxiety scale; MADRS; Mobility Inventory for Agoraphobia; OCI; Panic Rating Scale; Penn State Worry; PHQ-9; POMS; SPIN; Y-BOCS
	Bipolar disorder	Internal State Scale
	Eating disorders	EDE-Q
	Personality disorders	BPDSI; Zanarini scale for Borderline
	Post-traumatic stress disorder	Impact of Events Scale
	Psychotic symptoms	Auditory Hallucination Rating Scale; CAARMS; PANSS; Psychotic Symptom Rating Scales: Delusions
Physical symptoms	Adverse effects	Abnormal Involuntary Movement Scale
	Fatigue	Chronic Fatigue Questionnaire
Social functioning and overall well-being	Employment	Work and Social Adjustment Scale
	Quality of life, social functioning, well-being	EQ-5D; PSYCHLOPS; SF-36; The How are you scale; Life Satisfaction Index
	Social functioning & functioning disabilities	RFS; SDS; Schwartz Outcome Scale; Social Adaptation Self-Evaluation Scale; Social Adjustment Scale; Social Functioning Scale; Social Summary Rating Scale
Services: settings, evaluation, organisation of care and related patient views	Forensic	Historical Clinical Risk 20
	Healthcare and needs assessment	CANSAS; CUES; FACE; HoNOS; TAG; Maslach Burnout Inventory
	Patient perceptions of care	IPQ; Client Satisfaction Questionnaire; PEQ Part 1; PEQ Part 2
	Outcome of psychological therapies	CORE (several measures); Outcome Rating Scale; Session Rating Scale; The Barrett Lennard Inventory; The Inventory of Interpersonal Problems
	Recovery (and interaction with services)	DREEM; Mental Health Recovery Star; Ohio consumer assessment I & II
	Service planning	WHO DAS-S

AUDIT: Alcohol Use Disorder Identification Test; BDI: Beck Depression Inventory; BPDSI: Borderline Personality Disorder Severity Index; CAARMS: Comprehensive Assessment of At Risk Mental States; CANSAS: Camberwell Assessment of Needs Short Appraisal Schedule; CORE: Clinical Outcomes in Routine Evaluation; CUES: Carers & Users Expectations of Service; DREEM: Developing Recovery Enhancing Environments Measure; EDE-Q: Eating Disorders Examination Questionnaire; EPDS: Edinburgh Postnatal Depression Scale; FACE: Functional Assessment of the Care Environment; GAD-7: Generalised Anxiety Disorder - 7 items; GHQ-12: General Health Questionnaire - 12 items; HADS: Hospital Anxiety and Depression Scale; HAI: Health Anxiety Inventory; HoNOS: Health of the Nation Outcomes Scales; IPQ: Illness Perception Questionnaire; MADRS: Montgomery & Asberg Depression rating tool; OCI: Obsessive Compulsive Inventory; PANSS: Positive And Negative Syndrome Scale; PEQ: Patient Experience Questionnaire; PHQ-9: Patient Health Questionnaire - 9 items; POMS: Profile of Mood States; PSYCLOPS: Psychological Outcome Profiles; RFS: Role Functioning Scale; SDS: Sheehan Disability Scale; SPIN: Social Phobia Inventory; TAG: Threshold Assessment Grid; WHO DAS-S: World Health Organization Short Disability Assessment Schedule; Y-BOCS: Yale-Brown Obsessive-Compulsive Scale

Jacobs, 2009

The objective of the study conducted by Jacobs (2009) was to identify the most commonly used outcome measures in the UK mental health services that would be suitable to convert into a PBM. The author used the following criteria in order to identify suitable measures for this purpose:

- Ability to capture a wide range of mental health problems, so it is possible to use as generic mental health CSMs
- Wide (national) coverage in British NHS services
- Applicability in a number of care settings
- Routine collection in clinical practice
- High level of linking to activity data
- Feasibility of conversion into a PBM
- Availability of time series data

Coverage in NHS services was assessed based on the results of a systematic literature review and interviews with policymakers, academics and NHS staff involved in outcome measurement in adult mental health services. Based on the above criteria, two measures were identified as good candidates for translation into a utility index: **HoNOS** and the Clinical Outcomes in Routine Evaluation – Outcome Measure (**CORE-OM**). The study findings indicated that both HoNOS and CORE-OM were used in routine clinical practice and were probably the measures with the widest coverage in the NHS at the time the study was conducted, even though this coverage was somewhat patchy and in some areas non-existent. Regarding applicability across different health settings, the author acknowledged that this was difficult to achieve as most instruments considered appropriate in one setting might be inappropriate in another. HoNOS, which is a clinician-rated measure, was found to be mainly used in secondary care settings for patients with severe and enduring mental illness, while CORE-OM, which is patient-reported, covered patients that were primarily treated in the community setting or received psychological therapy (mostly people with depression and anxiety disorders). Time series data on activity and outcome were available for both HoNOS and CORE-OM, but it was reported that data quality was a concern and access to CORE-OM would

need to be negotiated. The author expressed the opinion that the valuation of either HoNOS or CORE-OM in order to derive a PBM would be an ‘extremely complex’ task.

Finally, the aim of the report by Hampson and colleagues (2011) was to provide guidance on the use of outcome measures in mental health based on what is of clinical value to patients and clinicians and what is feasible in practice. The report suggested a list of CSMs as a guide to clinicians and patients, which were selected from measures that had been shortlisted in the Mental Health Outcomes Compendium (National Institute for Mental Health in England, 2008).

2.4.3 Other studies evaluating outcome measurement in mental health (not UK-focused)

The review by Flynn (2002) was the first in a series of systematic reviews of available standardised mental health outcome measures to identify those most suitable for use by the US Veterans Health Administration (VHA) mental health services. The review focused on generic measures of mental health status, psychiatric symptoms and functioning that could be used in conjunction with disease-specific ones to monitor treatment effectiveness for clinical planning. Suitability of the measures in all reviews of the series was determined by a number of selection criteria, including:

- congruence of the original purpose of the measure with VHA intended use
- ability of the measure to capture multiple aspects of disease including both symptoms and functioning (‘multidimensionality’)
- applicability to serious mental illness (such as schizophrenia and major depression)
- reliability
- validity
- responsiveness to change
- feasibility for routine use (i.e. imposing minimal burden to clinicians and patients)

- interpretability by non-professionals
- availability in electronic form (for entry and analysis)
- reasonable cost

Based on the above criteria, the author shortlisted 5 generic mental health measures that met all criteria and were recommended as most appropriate for use by the US VHA mental health services, that is, the Behaviour and Symptom Identification scale-12 item (**BASIS-12**), **BPRS**, the Compass Out-Patient (**Compass-OP**), **GAF** and **HoNOS**, Two further measures, the Camberwell Assessment of Need (**CAN**) and **TAG** met almost all criteria, missing only the criterion for availability in electronic form, and thus were also considered for use by the US VAH mental health services.

Flynn (2004) reviewed existing compendia and reviews to inform the US VAH mental health services on outcome measures most appropriate for depression, using the same criteria listed above. The author identified 15 appropriate measures, 10 specific to depression and 5 generic ones. Of the 5 generic measures, 2 were not specific to mental health; these were the SF-36 and the QWB, including the self-administered version (QWB-SA). The other 3 were generic mental health measures and included **GAF**, the Mental Health Inventory (**MHI**) and the Sheehan Disability Scale (SDS); the latter is a tool focusing on functioning status in the areas of work/school, social life and family life in people with mental disorders (mainly depression and anxiety), and therefore is considered too narrow in scope to form the basis of a generic mental health PBM.

Finally, Flynn (2003) reviewed existing compendia and reviews to inform the US VAH mental health services on outcome measures most appropriate for schizophrenia, using the same set criteria described for the previous two reports. The author identified 13 appropriate measures, of which 10 are specific to schizophrenia and 3 are generic mental health measures. The latter included the Clinical Global Impression scale (**CGI**), the Role Functioning Scale (**RFS**), and **BPRS**.

Burns and Patrick (2007) conducted a systematic literature review to identify outcome measures used most frequently to assess social functioning in schizophrenia, and to assess their psychometric properties. Of the 3 measures that were identified as the most widely used for this purpose, 2 were measures of functioning used in general psychiatry, that is, **GAF** and **GAS**. GAF was identified as the most widely used social functioning scale for people with schizophrenia, providing a reliable assessment of psychological, social and occupational functioning. GAS is precursor of GAF.

McCabe and colleagues (2007) conducted a non-systematic review to identify PROMs in schizophrenia and to assess their psychometric properties. The authors identified 20 measures in total. Of these, 6 focused on symptoms and needs assessment, 9 assessed the clinician-patient therapeutic relationship, the patients' attitude toward therapy and their satisfaction with services, and another 5 aimed to capture the psychological well-being of patients. Of the 6 measures focusing on symptoms and needs assessment, 4 were generic mental health measures that can be used across different mental disorders; these included the Brief Symptom Inventory (**BSI**), the **SCL-90-R, CAN** and **CANSAS**. BSI was reported to have high internal consistency, test-retest reliability, and construct (convergent and known groups) validity. Evidence for SCL-90-R indicated high internal consistency and adequate test-retest reliability. Both CAN and its shortened version, CANSAS, were found to have high face validity and reliability. The 9 measures assessing patients' views and attitude toward therapy and services have a narrow scope and were deemed not appropriate to form the basis for a generic mental health PBM; therefore these are not discussed further in this chapter. Finally, all 5 PROMs of psychological well-being were generic measures that can be used across different mental disorders; these included the Empowerment Scale, the Self-Esteem Scale, the Sense of Coherence Scale (SOC), the Mental Health Recovery Measure (MHRM) and the Recovery Assessment Scale (RAS). These measures have a narrow scope, focusing on specific aspects of psychological well-being rather than capturing a range of aspects that constitute a person's HRQoL and thus were not regarded suitable candidates for the derivation of a PBM.

2.4.4 Identification of appropriate outcome measures as candidates for the derivation of a mental health specific preference-based measure

The systematic review of existing reviews revealed that there is a range of validated CSMs that can be used as generic measures across different mental disorders, have quite a wide scope (i.e. they are not limited to capturing a specific aspect of HRQoL), and therefore could potentially form the basis for a generic mental health PBM. Among the most widely used and/or recommended measures were the BPRS, CAN and its shorter form CANSAS, the CORE measurement tools, GAF and its precursor GAS, HoNOS and TAG.

BPRS is a clinician-rated scale designed to measure major psychotic and non-psychotic symptoms (Overall & Gorham, 1962); it is mainly used in patients with schizophrenia.

CAN (and its shortened form, CANSAS) is a clinician-rated measure aiming to assess the needs of people with severe mental illness (Phelan et al., 1995).

The CORE measurement tools, which consist of CORE-OM (Evans et al., 2000) and a number of other inter-dependent measures that have been developed around the CORE-OM, are PROMs designed to measure aspects of psychological distress, including relevant symptoms and well-being, before and after therapy, thus providing a routine outcome measurement system for psychological therapies and some areas of psychiatry.

GAF (and its precursor, GAS) is a clinician-rated scale that evaluates patients' psychological, social and occupational functioning covering a range from positive mental health to severe psychopathology (Jones et al., 1995).

HoNOS is a clinician-rated measure designed to measure the health status and social functioning of people with severe mental illness (Wing et al., 1998).

TAG is also clinician-rated, and is designed to evaluate the severity of symptoms in people with mental disorders, so as to prioritise those in need for specialist mental health care (Slade et al., 2000).

In selecting an appropriate measure of those briefly described above as the basis for the derivation of a generic mental health PBM, emphasis was given to reviews that had a special focus on the UK mental health research and/or practice, as these examined factors such as coverage, routine collection and views of psychiatrists and patients in the British NHS context.

The most relevant of the reviews considered was the one by Jacobs (2009), as the aim of that review reflected the objective of the review of reviews conducted for this thesis, i.e. it aimed to identify the most commonly used outcome measures in the UK mental health services that would be suitable to convert into a PBM. The author identified HoNOS and CORE-OM as the two most suitable measures based on their ability to capture a wide range of mental health problems, coverage and routine collection in NHS services, availability of activity and time series data, and applicability in various care settings.

In addition to the recommendation by Jacobs (2009), HoNOS was shortlisted in the Mental Health Outcomes Compendium for receiving a high quality scoring and being recommended by stakeholders (National Institute for Mental Health in England, 2008). It was also identified as one of the most widely used outcome measures in psychiatric research and practice by Gilbody and colleagues (2003). Finally, with regard to US guidelines, it was one of the generic mental health measures recommended by Flynn (2002) for use by VAH mental health services.

The CORE measurement tools (which include the CORE-OM) were also shortlisted in the Mental Health Outcomes Compendium for receiving a high quality scoring and being recommended by stakeholders (National Institute for Mental Health in England, 2008). The CORE-OM was not identified as a widely used measure in the review by Gilbody and colleagues (2003), but this is

possibly attributable to the fact that CORE-OM was not developed until 1998, and this review covered the years 1956-2000, while the interviews with UK psychiatrists must have taken place before 2002, when CORE-OM use was likely not widespread yet. CORE-OM was not among the recommended measures in the non-UK focused reviews, most likely because either a. these considered evidence available up to early 2000s, when CORE-OM was not fully validated and/or widely used or b. these focused on outcome measures for patients with schizophrenia, whereas CORE-OM has not been designed for use in this patient population.

2.5 Selection of an appropriate outcome measure as the basis for the derivation of a generic mental health-specific preference-based measure

This section reviews in more detail the properties, usage and applications of HoNOS and the CORE-OM, which appeared to be the leading candidates for the derivation of a generic mental health PBM following the findings of the review of reviews. The aim of this section is to justify the selection of CORE-OM as the most appropriate between the two measures for the derivation of a new PBM specific to mental health.

2.5.1 The Health of the Nation Outcome Scales (HoNOS)

HoNOS is a clinician-rated questionnaire developed by the Royal College of Psychiatrists' Research Unit in consultation with clinical experts, following its commissioning by the UK Department of Health in 1993 "to develop scales to measure the health and social functioning of people with severe mental illness".[1] HoNOS was developed in 4 phases (Wing et al., 1998). Phase I comprised a literature review of existing measures and the development of a draft measure after consultation with clinical experts. In Phase II, the drafted measure HoNOS-I was shortened to version HoNOS-II, following pilot tests on simplicity in structure, acceptability to clinicians, and sensitivity to change. HoNOS-II was tested against using the same criteria and was modified to version HoNOS-III, which was subsequently tested in field trials comprising Phase III of the project. These larger scale trials assessed the properties that

[1] http://www.rcpsych.ac.uk/crtu/healthofthenation.aspx [Accessed 22 April 2013].

were previously tested in Phase II, plus the reliability of the measure and the ability of HoNOS-III ratings to describe distinct clinical profiles for different diagnostic groups. Results of these trials led to further amendments and the development of the final scale, the HoNOS, which was then re-tested in Phase IV on all previous attributes plus its comparability to existing validated larger measures.

HoNOS includes 12 items, each with 5 levels of response: ‘no problem’, ‘minor problem requiring no action’, ‘mild problem but definitely present’, ‘moderately severe problem’ and ‘severe to very severe problem’. The 12 items cover 4 domains: ‘behaviour’, ‘impairment’, ‘symptoms’ and ‘social functioning’ (Wing et al., 1998). Depending on the level of response, each item can get a score from 0 (‘no problem’) to 4 (‘severe to very severe problem’). The sums of item ratings represent a clinical judgement of severity of the mental disorder: for example, the more 0s the lower the severity, the more 4s the greater the severity. Item scores in each domain can also be added to give a total subscale score. Changes in subscale scores between two time points, most typically between the start and the end of an episode of care, provide an indication of change in the patient’s health status specific to each domain. A total scale score can be derived by adding all item scores and can get a value between 0 (best possible score) and 48 (worst possible score). However, the developers advise against estimation of a total scale score, because the 12 items are so wide in their coverage that significant improvements in one domain may be cancelled out by deterioration in another, thus potentially giving the wrong impression that no improvement has occurred within a time period, e.g. over a completed episode of care. The domain structure of HoNOS is shown in Table 5.

Table 5. The domain structure of HoNOS

Domain	Item
Behaviour	1. Overactive, aggressive, disruptive or agitated behaviour 2. Non-accidental self-injury 3. Problem-drinking or drug-taking
Impairment	4. Cognitive problems 5. Physical illness or disability problems
Symptoms	6. Problems associated with hallucinations and delusions 7. Problems with depressed mood 8. Other mental and behavioural problems
Social functioning	9. Problems with relationships 10. Problems with activities of daily living 11. Problems with living conditions 12. Problems with occupation and activities

HoNOS versions are available for children and adolescents (Gowers et al., 1999), older people (Burns et al., 1999), people with learning disabilities (Roy et al., 2002), people in forensic services (Dickens et al., 2007), and people with acquired brain injury (Fleminger et al., 2005). Copyright of HoNOS is owned by the Royal College of Psychiatrists. The College allows the free use, copy and reproduction of HoNOS score-sheets for use in NHS-funded care without requested permission. However, commercial copying, renting and adaptation are prohibited.

Overall, HoNOS has been shown to be a valid, reliable and responsive measure, acceptable to clinicians and appropriate for routine outcome measurement (Amin et al., 1999; Andreas et al., 2010; Eagar et al., 2005; Kisely et al., 2007 & 2010; Kodagalli et al., 2012; McClelland et al., 2000; Oiesvold et al., 2011; Orrell et al., 1999; Page et al., 2001; Pirkis et al., 2005), although concerns regarding its reliability and sensitivity have been expressed by a number of researchers, questioning its usefulness as a routine outcome measure in mental health services (Audin et al., 2001; Bebbington et al., 1999; Brooks, 2000; Duke, 2010; Orrell et al., 1999; Sharma et al., 1999; Slade et al., 1999; Trauer et al., 1999). HoNOS correlates moderately with other widely used validated measures such as the EQ-5D index (Kodagalli et al., 2012), CANSAS (Slade et al., 1999), and CORE-OM (Leach et al., 2005), but its correlation with other measures such as the mental component score of SF-36

(Brooks, 2000) and the SCL-90-R (Brooks, 2000; Oiesvold et al., 2011) has been found to be weak.

Over the years, HoNOS has been recommended as the main outcome measure for use in mental health services by a number of advisory bodies and strategic plans including the English National Service Framework for Mental Health (Department of Health, 1999), the working group to the Department of Health on outcome indicators for severe mental illnesses (Charlwood et al., 1999), and the Outcomes Reference Group, a group established by the Department of Health in 2002 to advise on best practice guidance (Fonagy et al., 2004). More recently, the strategic plan on mental health designed by the coalition government 'No health without mental health' recommends the use of HoNOS for measuring outcome in people with severe mental illness and acknowledges the measure's widespread use and acceptability (HM Government & Department of Health, 2011). HoNOS is one of the recommended quality and outcomes indicators to be used as part of the introduction of the Payment by Results currencies and local tariffs for mental health services, and the only indicator of those recommended that is routinely collected across NHS services (Quality and Outcomes Sub Group of the Product Review Group for Mental Health Payment by Results, 2011). It needs to be noted, though, that, according to its developers, HoNOS has not been designed for use in primary care; consequently, it may not be appropriate for outcome measurement in people with mild or moderate mental disorders that are treatable in primary care settings.

Further to the lack of applicability of HoNOS in primary care, its main disadvantage in being used as the basis for a generic mental health PBM is the fact that it is clinician-rated and not patient-reported. PBMs like EQ-5D, SF-6D and HUI-3 have been traditionally patient-reported, i.e. they collect information on HRQoL from patients themselves (rather than clinicians); these HRQoL ratings have been subsequently linked to utility values expressing the preferences of members of the general population. In this sense, PBMs can be regarded as a special form of PROMs (Stevens & Palfreyman, 2012). To this direction, NICE explicitly requires measurement of HRQoL changes be elicited

from patients, and, if this is not possible, from their carers, rather than healthcare professionals (National Institute for Health and Care Excellence, 2013). Consequently, the clinician-rated HoNOS was not considered appropriate for the derivation of a generic mental health PBM.

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

The CORE-OM is a PROM that was developed by a multicentre collaborative group as a result of winning a competitive tender to develop an outcome measure at a conference for the Mental Health Foundation in 1993.[2] The rationale for developing a new measure was the "*need for a pragmatic, user-friendly measure that taps pan-theoretical 'core' components of patients' distress*" that would be "*implemented on a broad basis across adult mental health services in order to enable benchmarking and standardise outcome*" (Barkham et al., 1998). Development of CORE-OM was carried out in 4 phases (Barkham et al., 2001). Phase I involved a survey of the views of providers and purchasers of mental health services about the current use of outcome measures in the services and the desirable aspects of a new outcome measure. In Phase II, 6 independent groups of raters assessed the results of the survey undertaken in Phase I. In Phase III the development team designed the criteria that the new measure should meet and drafted the items of the new measure based on the results of the previous 2 phases. The drafted items were then tested for qualitative feedback from a wider group of more than 40 therapists, researchers and lay people. The CORE-OM was finalised in Phase IV of the project.

The CORE-OM is a measure of psychological distress that was designed to evaluate the effectiveness of psychological therapies across multidisciplinary services in the UK (Barkham et al., 2001; Evans et al., 2000). 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 domains considered by practitioners to be necessary components in a 'core' measure: 'subjective well-being' (4 items), 'problems' (4 items on depression, 4 items on

[2] http://www.coreims.co.uk/index.html [Accessed 25 April 2013].

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. Depending on the level of response, each item is scored from 0 (‘not at all’) to 4 (‘most or all the time’), with the exception of positively worded items, the scores of which are reversed. The CORE-OM clinical score is then calculated by adding all 34-item scores, multiplying by 10 and dividing by 34. The CORE-OM clinical score can get values between 0-40, with 10 being considered the cut-off point between clinical and non-clinical cases. A clinical score 10 to <15 indicates mild psychological distress, 15 to <20 moderate distress, 20 to <25 moderate to severe distress, and 25 to 40 severe psychological distress (Barkham et al., 2006). The 34 items of CORE-OM categorised by domain and sub-domain are presented in Table 6.

Table 6. The conceptual domain structure of CORE-OM

Domain	Item N^{o}	Item description
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've 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 someone 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

In addition to CORE-OM, other CORE system products are available for outcome measurement (Gray & Mellor-Clark, 2007). These include the short form A (CORE - SFA) and short form B (CORE - SFB) which consist of 18 items each and are usually used in research studies at alternate sessions instead of the CORE-OM to reduce memory effects (Cahill et al., 2006), the brief form CORE-10 used in routine practice for initial screening and session-by-session monitoring (Barkham et al., 2013), the even shorter CORE-5 used also in routine practice for session-by-session monitoring, the CORE-GP for measuring the mental health of the general or student population (Sinclair et al., 2005), the YP-CORE appropriate for young people (Twigg et al., 2009), the CORE-LD for use in people with learning disabilities (Brooks et al., 2013; Marshall & Willoughby-Booth, 2007), the CORE Goal Attainment Form for tracking goal attainment (Proctor & Hargate, 2013), and the ARM-5, which is a measure of therapeutic alliance. Further to these, the CORE System includes also a therapy assessment form (CORE-A) (Barkham et al., 2005b) and an end of therapy form (Connell et al., 2006), which have been adapted for use with young persons, at workplace, and in further and higher education. The CORE system is supported by special software (CORE-PC and CORE-NET), as well as training and backup services provided by the CORE Information Management Systems (CORE IMS).

The CORE System Trust, a not-for-profit company, holds the copyright of CORE measures. CORE measures may be photocopied freely provided that they are not modified or used for financial gain. However, creating electronic versions for inclusion in software systems other than those provided by CORE IMS requires written permission from the Trustees of the CORE System Trust.

CORE-OM comprises a valid, reliable, responsive and acceptable effectiveness measure across a wide range of practice settings offering psychological therapies, including primary and secondary care (Barkham et al., 2001 & 2005b; Evans et al., 2002 & 2003). It has been validated in older populations (Barkham et al., 2005a). Its diagnostic value for depression is as good as clinician-rated measures (Gilbody et al., 2007) and its correlation with other widely used CSMs such as the BDI (Cahill et al., 2006; Leach et al.,

2006), the HoNOS (Leach et al., 2005) the HAM-D (Cahill et al., 2006), and the Clinical Interview Schedule - Revised (CIS-R) (Connell et al., 2007) is moderate to high. CORE-OM has been administered to both general population and clinical samples in the UK to assess normative values and find appropriate cut-offs between clinical and non-clinical cases; this has led to the construction of normative tables with distinct severity levels for clinical and non-clinical UK population (Barkham et al., 2006; Connell et al., 2007; Evans et al., 2002). CORE System outcome measurements have also been used to establish benchmarks against which services can review their own data, contributing to service assessment and improvement in the quality of care (Barkham et al., 2001; Evans et al., 2003; Mellor-Clark et al., 2006; Mullin et al., 2006).

Jacobs (2009) reported that CORE-OM was considered the most widely used outcome measure in psychological therapy and counselling services in the UK, with the CORE IMS database covering around 100,000 patients per annum. According to the CORE System website, in 2010 CORE-OM software was used by over 250 organisations, including 40 primary care services, 40 secondary and tertiary care services, 30 workplace services, 80 voluntary sector services, 30 university and 10 private services.

The Improving Access to Psychological Therapies (IAPT) programme was initiated in 2006, aiming to support the British NHS in delivering evidence-based psychological therapies for people with depression and anxiety disorders, ensuring wide and timely access to services and treatments, improvement in service users' health and well-being, employment, benefit, and social inclusion status, as well as increased patient choice and high levels of satisfaction. The current IAPT data handbook (IAPT, 2011) recommends the use of 4 measures on all patients seen in IAPT at minimum, based on their suitability, free access and wide use. Of these, 3 are specific to depression (PHQ-9), generalised anxiety (Generalised Anxiety Disorder - 7 items – GAD-7) and phobias (IAPT Phobia Scales), and one is a social functioning measure (Worker and Social Adjustment Scale - WSAS). Nevertheless, the 2008/2009 IAPT toolkit (IAPT, 2008) acknowledged CORE-OM as a measure widely used

to monitor changes in psychological health and well-being, that covered a wider range of client-presenting problems than the disorder-specific measures PHQ-9 and GAD-7. Although the CORE-OM was not part of the minimum IAPT dataset, the IAPT toolkit encouraged sites already using CORE-OM to continue its use pre- and post-treatment in addition to measures recommended by IAPT. Sites that continued use of the CORE-OM were advised to supplement its use with shorter versions, such as the CORE-10, in order to save time and record outcomes and recovery data for as many patients as possible.

Compared with HoNOS, the CORE-OM has wider applicability across service settings, as it has been validated for use in both primary and secondary care. More importantly, in contrast to HoNOS, it is a PROM and therefore it is more suitable than HoNOS to form the basis for the derivation of a new PBM.

One disadvantage of the CORE-OM is that by design it is suitable for use in people with common mental health problems; these include various types of depression and anxiety, including unipolar depression, GAD, mixed anxiety and depressive disorder, phobias, OCD and panic disorder. CORE-OM has not been designed for use in people with severe mental illness such as schizophrenia, bipolar and personality disorders, and therefore it may not be appropriate to use in such populations. Consequently, CORE-OM cannot form the basis for a generic mental health PBM that can be used across all mental disorders. On the other hand, the review of reviews did not identify any other CSM that has all the advantages of CORE-OM (patient-reported, broad coverage of symptoms and aspects of HRQoL, valid and responsive, wide use within the NHS, applicability to primary and secondary care settings) *and* can be used as a generic mental health measure across the full range of mental disease. In any case, common mental health problems alone have a prevalence that reaches 18% in people aged 16-64 years living in England (for comparison, psychotic disorders are prevalent in only 0.4% of this population) (McManus et al., 2009). Therefore, CORE-OM is applicable to the large majority of people with mental disorders, including those with common mental health problems and potentially to a range of populations with more severe

mental illness. For this reason, and considering its advantages compared with other candidate CSMs, CORE-OM was selected for the derivation of a mental health-specific PBM for people with common mental health problems.

Given that the CORE System includes forms that are more concise than the CORE-OM, such as CORE-SFA and CORE-SFB, and, in particular, CORE-10 and CORE-5, there was the question of whether to derive the PBM from these shorter forms, or, indeed, whether these shorter forms could directly form a PBM, following a valuation survey. After reviewing these shorter forms, the purpose of their development and the methods employed for their construction, it was decided that it was preferable to assess the full pool of the 34 CORE items for their appropriateness and suitability for inclusion in a PBM, rather than to limit the pool of items or to use existing brief forms of the CORE-OM that were developed for different purposes. Furthermore, it was felt that the new PBM should be derived from the original CORE-OM by adopting and/or adapting validated approaches described in the literature for this purpose.

In addition to the review of its psychometric properties, its coverage and applicability across different service settings, the appropriateness of CORE-OM to form the basis for the derivation of a generic mental health PBM was examined by assessment of its content validity by matching its items against the 7 themes identified as having the greatest impact on HRQoL in people with mental disorders (Brazier et al., 2014; Connell et al., 2012). Each item was matched to one ‘primary’ HRQoL domain, although some items could be potentially matched to more than one domain. Table 7 shows the results of this assessment.

Table 7. Content validation of CORE-OM against the main domains of health-related quality of life that are important to people with mental disorders [as identified by Brazier and colleagues (2014) and Connell and colleagues (2012)]

Domain	Sub-theme and summary description	CORE-OM item
Subjective well-being & ill-being	Distress; associated with depression, experience of psychosis and mania and anxiety Depressive mood; associated with poor concentration, low energy and poor motivation Fear or panic and anxiety; can be caused by stressful social situations Psychosis-related distress; caused by critical voices, difficult to differentiate from reality Positive well-being: happiness and enjoyment; feeling peaceful, calm, relaxed and safe Energy and motivation (lack of both often caused by lack of sleep)	2 I have felt tense, anxious or nervous 5 I have felt totally lacking in energy and enthusiasm 13 I have been disturbed by unwanted thoughts and feelings 14 I have felt like crying 15 I have felt panic or terror 17 I have felt overwhelmed by my problems 18 I have had difficulty of getting to sleep or staying asleep 20 My problems have been impossible to put to one side 27 I have felt unhappy 28 Unwanted images or memories have been distressing me
Activity & functioning	Positive: work, hobbies or social interaction Negative: stressful if too demanding; fear of stress may result in avoiding enjoyable activities	11 Tension/anxiety have prevented me doing important things 32 I have achieved the things I wanted to
Social well-being, belonging & relationships	Relationships: close friends and family Social relationships Reactions of others – understanding, acceptance and stigma Sense of belonging	1 I have felt terribly alone and isolated 3 I have felt I have somebody to turn to for support when needed 6 I have been physically violent to others 10 Talking to people has felt too much for me 19 I have felt warmth or affection for someone 22 I have threatened or intimidated another person 25 I have felt criticised by other people 26 I have thought I have no friends 29 I have been irritable when with other people 33 I have felt humiliated or shamed by other people
Self-perception	Self-identity Self-efficacy, self-esteem and self-acceptance Self-stigma	4 I have felt ok about myself 9 I have thought of hurting myself 12 I have been happy with the things I've done 30 I have thought I am to blame for my problems & difficulties 34 I have hurt myself physically or taken risks with my health
Control, autonomy & choice	Dependence and independence – relating to support Self-control: mainly related to relief / management of symptoms, usually through medication Choice: money and access to resources	7 I have felt able to cope when things go wrong 21 I have been able to do most things I needed to
Hope & hopelessness	Dreams and goals, involvement in activities that give meaning and purpose Hopelessness; lowering of aspirations	16 I made plans to end my life 23 I have felt despairing or hopeless 24 I have thought it would be better if I were dead 31 I have felt optimistic about my future
Physical health	Physical comorbidity or experience associated with mental health problem	8 I have been troubled by aches, pains, physical problems

CORE-OM appears to cover all major domains of HRQoL that are important to people with mental disorders. Several of its items express subjective well-being and capture symptoms of anxiety and depression that are relevant to ill-being. One of its anxiety items (11. Tension/anxiety have prevented me doing important things) covers negative aspects of activities & functioning, while another item (32. I have achieved the things I wanted to) is implicitly related to positive aspects of this domain. CORE-OM has several items capturing close and social relationships and belonging. Risk-to-other items can also be considered relevant to this theme. Regarding self-perception, CORE-OM has a number of items relating to self-esteem, whereas the self-harming items can be regarded as indicative of lack of self-esteem and self-acceptance. CORE-OM does not directly capture control, autonomy and choice, although two of its items (7. I have been able to cope when things go wrong and 21. I have been able to do most things I needed to) suggest autonomy and control over life. CORE-OM includes items with explicit reference to feelings of hope/optimism and hopelessness; items expressing thoughts of suicide are also indicative of feelings of hopelessness. Finally, item 8 of CORE-OM (I have been troubled by aches, pains, physical problems) captures the theme of physical health.

The CORE Outcome Measure form is provided in Appendix 3.

2.5.3 Conclusion

Following a systematic review of published reviews of outcome measurement in mental health, HoNOS and CORE-OM were identified as the leading candidates for the derivation of a generic mental health PBM.

HoNOS is a measure of severe and enduring mental illness. Its psychometric properties have been tested and generally it is considered a valid, reliable and responsive outcome measure. It can be used for free within the British NHS, and in fact is a widely used measure in the UK clinical practice that has been advocated by many advisory bodies for routine outcome measurement. HoNOS has not been designed for use in primary care, and therefore may not be appropriate for use in people with mild and moderate mental illness. HoNOS is clinician-rated, whereas PBMS are traditionally patient-reported, and this was considered a major disadvantage against its use as the basis for the

derivation of a generic mental health PBM. Therefore, HoNOS was excluded from further consideration.

On the other hand, the CORE-OM was selected as the basis for the derivation of a new mental health-specific PBM based on its following properties and characteristics:

- Broad coverage of symptoms and aspects of HRQoL, including both mental and physical health aspects
- Psychometric properties: established construct validity, responsiveness, reliability and acceptability
- Wide coverage within the British NHS
- Applicability across primary and secondary settings
- Free use
- Being patient-reported
- Representation of the areas of HRQoL that have been identified to be important in people with mental disorders (content validity)
- Appropriate for outcome measurement in people with common mental health problems, which are the most prevalent mental disorders in the UK

2.6 Overall conclusion

The findings of the systematic review of reviews on the performance of generic PBMs in people with mental disorders seem to justify the concerns that have been expressed regarding the appropriateness of generic PBM use in the area of mental health. The limited available evidence indicates that generic PBMs perform satisfactorily in depression, but less so in anxiety and personality disorders. The picture is mixed in schizophrenia and bipolar disorder. Qualitative evidence suggests that generic PBMs fail to capture aspects of HRQoL that are important to people with mental disorders.

The systematic review of reviews of outcome measurement in mental health revealed that there is a breadth of validated measures in this area, which vary in focus (psychopathology/symptoms versus impact on patients' lives/HRQoL

aspects), scope (global versus disorder- or domain-specific), purpose (assessment of symptom severity versus assessment of patient needs) and intended rating population (patient-reported versus clinician-rated). Based on its psychometric properties, broad coverage of a range of symptoms and aspects of HRQoL that are relevant to mental health populations, wide and free usage within the British NHS, applicability across primary and secondary settings and the fact that it is patient-reported, CORE-OM was selected as the basis for the derivation of a mental health-specific PBM that is relevant to people with common mental health problems.

The next chapter reviews the methods reported in the literature for the derivation of health state classifications amenable to valuation from existing longer measures. Chapters 4, 5 and 6 report the methods adopted and the process that was followed in this thesis in order to derive a new PBM that is relevant to people with common mental health problems from the CORE-OM.

Chapter 3. Methods for deriving health state descriptions from existing longer outcome measures – systematic literature review

3.1 Introduction

As reported in Chapter 2, following a review of the properties of generic PBMs in the area of mental health and confirmation of their inadequacy to capture relevant aspects of HRQoL in people with mental health problems, the CORE-OM was selected for the derivation of a new PBM that is relevant to people with common mental health problems. Development of a PBM is a 3-step process that involves the description of health states usually by a health state classification system, the valuation of a selection of health states in a valuation survey, and further econometric modelling that allows attaching an appropriate utility value to every health state described by the health state classification. It has been suggested that respondents can receive, process and remember about seven pieces of information plus or minus two, depending on the complexity of the statements (Miller, 1956). Therefore, health state classifications amenable to valuation need to be concise, comprising a manageable number of items and response levels; at the same time, they must be comprehensive enough to capture a range of relevant aspects and levels of HRQoL.

The CORE-OM consists of 34 items with 5 levels of response each that cover 4 major conceptual domains. Inclusion of all items of the CORE-OM in the health state classification system of the new PBM would result in the description of a massive number of potential health states that would be impractical to use and complicated to value in a valuation survey. As discussed in Chapter 1 (section 1.3.3), a concise health state descriptive system can be derived from a long measure such as the CORE-OM by selecting appropriate domains, items and levels. The selection process needs to identify the most representative domains and items of the CORE-OM to ensure that the new

health state classification retains to an acceptable degree the properties of the original measure and is characterised by minimum loss of information relative to the CORE-OM (Brazier et al., 2007).

The aim of this chapter is to systematically review methods that have been reported in the literature for the derivation of health state descriptive systems amenable to valuation from existing longer outcome measures, in order to identify and adopt or adapt appropriate methods that could be used for the derivation of a health state classification from the CORE-OM. A systematic search of the literature was undertaken for this purpose. This chapter provides an overview of the methods and the results of the systematic literature search and subsequently describes and critically reviews the methods proposed in the literature for the derivation of health state descriptive systems from existing, non-preference-based outcome measures.

3.2 Systematic search of the literature: methods and overview of results

The systematic search of the literature aimed to identify studies reporting methods for the derivation of health state descriptions amenable to valuation from existing outcome measures. The following databases were searched from inception for this purpose:

Via OVID interface

1. EMBASE (1980 to current)
2. MEDLINE
3. PsycInfo
4. Health Management Information Consortium (HMIC)

Via Wiley interface

5. Cochrane Database of Systematic Reviews (CDSR)
6. Cochrane Methodology Register
7. Health Technology Assessment (HTA) database
8. Database of Abstracts of Reviews of Effects (DARE)

The systematic search was initially carried out in March 2007 and updated in December 2012, after completion of the development of the new PBM that was

the subject of this thesis, in order to explore and describe recent trends in the derivation of health state descriptions from existing outcome measures. The search strategy used is an adaptation of the strategy reported in Brazier and colleagues (2012), which was employed for the identification of PBMs derived from existing CSMs. The review undertaken for this thesis considered not only PBMs but also health state descriptions amenable to valuation that were derived from existing measures, regardless of whether valuation of health states had been subsequently undertaken or reported in the literature. Thus, in the search strategy constructed for this thesis extra terms relating to health state descriptions were added. Moreover, the review undertaken for this thesis was not confined to health state descriptions (or PBMs) derived from CSMs, as appropriate methods that could be adopted or adapted in order to derive a PBM from the CORE-OM might have been used in the literature for the derivation of health state descriptions from generic measures as well. The search strategy used for the systematic search of the literature is provided in Appendix 4.

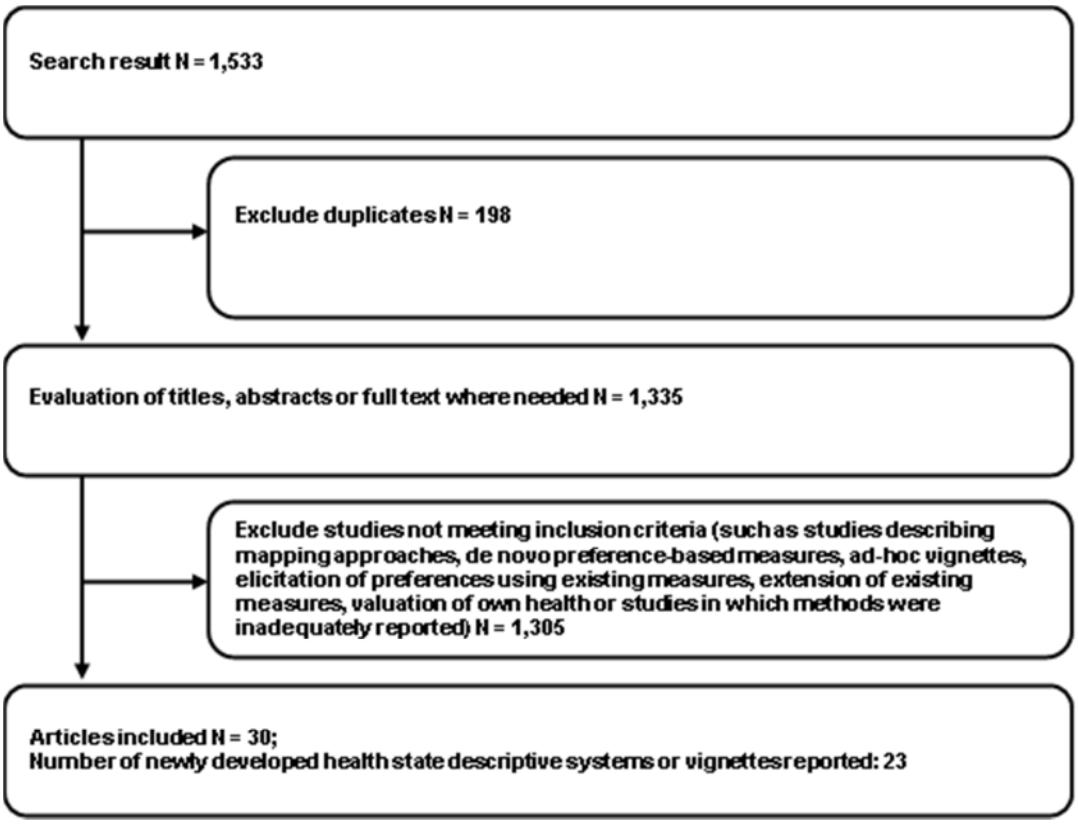
The following inclusion/exclusion criteria were applied to select studies identified by the search for further consideration:

- Papers were included provided that they described psychometric or other statistical methods for the derivation of health state descriptions amenable to valuation from existing measures (either generic or condition-specific). Papers reporting on vignettes derived from existing measures were also included in the review, if derivation of vignettes was based on statistical analysis of data, including psychometric methods.
- The purpose of the derivation of new health state descriptions should be their consideration in a valuation survey; studies reporting derivation of new measures from longer ones without aiming at developing health state descriptions to be used in valuation surveys were not considered.
- Measures derived from existing instruments where the selection of dimensions, items and levels was based on focus groups, expert opinion or on simple consideration of their relative ‘importance’ were not included in the review.

- Vignettes either developed *ad hoc* or described using items from existing instruments selected by focus groups were also not considered in the review
- Studies reporting *de novo* development of health state descriptive systems amenable to valuation were excluded
- Studies needed to provide adequate detail of the methods that would enable their adoption or adaption in order to derive a health state descriptive system from the CORE-OM
- Studies reporting mapping algorithms linking non-PBMs to existing generic PBMs were excluded
- Only papers published in English were considered
- No conference abstracts or poster presentations were considered in the review, as these did not describe methods in adequate detail

The systematic search identified 1,533 references in total. After excluding 198 duplicates, 1,335 titles and/or accompanying abstracts were screened for relevance against the set inclusion/exclusion criteria. Full texts of studies potentially meeting inclusion criteria (including those for which eligibility was not clear from the abstract) were obtained. After excluding studies clearly not relevant to the topic and studies not meeting inclusion criteria (for example, studies describing mapping approaches; *de novo* PBMs; *ad hoc* vignettes; elicitation of preferences using existing measures; extension of existing measures by adding extra dimensions and/or items; studies reporting regression analysis between patients' valuation of own health state and their responses to items of an existing measure; or studies in which methods were inadequately reported), 31 publications remained for inclusion in the review, describing 24 newly developed health state classifications or vignettes derived from existing measures using psychometric or other statistical methods (some newly constructed measures were described in more than one publications, each reporting on different stages of the new measure's development). A flow diagram showing the systematic process for selecting papers for the review is provided in Figure 5.

Figure 5. Flow diagram of the selection of publications in the systematic review of studies reporting methods for the derivation of health state descriptions amenable to valuation from existing outcome measures



A summary of the studies included in the systematic literature review, categorised according to the overall methodology used, is shown in Table 8. The systematic search update identified one publication relating to the content of this thesis, i.e. the derivation of a health state classification from the CORE-OM (Mavranezouli et al., 2011), as well as another 5 studies that used broadly the same methodology with that developed for this thesis [1 study by Young and colleagues (2010), 2 studies by Versteegh and colleagues (2012) and 1 study by Kowalski and colleagues (2012) who adopted the methods reported by Mavranezouli and colleagues (2011); and 1 study by Sundaram and colleagues (2009 & 2010) who proposed a similar methodology with that developed for this thesis independently and in parallel]. The methods used in these studies (which are shaded in grey in Table 8) are not reported in this chapter, since detailed description of the methodology developed for this thesis is provided in Chapter 4. Instead, these studies are briefly discussed in relation to the methodology developed for this thesis in Chapter 8 (section 8.2).

3.3 A critical review of the methods suggested in the literature for the derivation of health state descriptions from existing measures

The systematic literature review identified 2 broad approaches for the derivation of health state descriptions from existing measures. Both approaches rely on a combination of standard psychometric criteria and other statistical techniques, supplemented by expert opinion. The main approach that is widely reported in the literature is the construction of health state classifications that are typically multidimensional. An alternative approach for the derivation of health state descriptions from existing measures involves the development of plausible health state descriptions that cover a range of HRQoL levels in the form of vignettes. Finally, a hybrid approach that produces plausible health state descriptions from newly derived health state classifications has also been reported in the literature.

3.3.1 Derivation of health state classifications

Health state classifications are descriptive systems usually composed of a number of multilevel, single item dimensions that together can describe a universe of health states (Brazier et al., 2007). For example, EQ-5D (Brooks, 1996) has 5 items each covering a different dimension (mobility, self-care, usual activities, pain / discomfort, anxiety / depression); in the original form of EQ-5D, each item has 3 response levels (no problems - moderate problems - extreme problems). Consequently EQ-5D can describe $3^5 = 243$ different health states. Health state classifications contain a limited number of statements describing HRQoL and are therefore convenient to use; moreover, the number of resulting health states is manageable to value.

Concise health state classifications can be derived from existing measures using various statistical techniques in a 3-stage process that involves 1) assessment of the dimensionality of the original measure and selection of appropriate dimensions for the health state classification; 2) selection of items and item response levels for inclusion in the health state classification; and 3) validation of the new health state descriptive system (Brazier et al., 2012).

Table 8. Studies reporting the derivation of health state descriptions from existing measures

A. Development of new health state classifications			
i. Development of multidimensional health state classifications			
Primary methodology used	**Condition**	**Original measure (new measure)**	**Relevant references**
• FA / PCA for the identification of dimensions • ± Correlations between items and between items and original measure • Qualitative review of items and response levels for suitability and relevance – expert opinion for the selection / exclusion of items	General health	SF-36, SF-12 (SF-6D)	Brazier & Roberts, 2004; Brazier et al., 1998 & 2002
	Benign prostatic obstruction	IPSS	Kok et al., 2002
	Lung cancer	FACT-L	Kind & Macran, 2005; Lamers et al., 2007
• Classical psychometric criteria for the selection of items from each dimension/domain • ± IRT for selection of items • Expert opinion for the reduction of dimensions and/or items	Menopause	Un-named	Brazier et al., 2005b
	Urinary incontinence	KHQ	Brazier et al., 2008
	Sexual quality of life	SQOL (SQOL-3D)	Ratcliffe et al., 2009
	Paediatric atopic dermatitis	Un-named	Stevens et al., 2005
• PCA for the establishment of dimensions/domains • Rasch analysis for the selection/exclusion of items in each dimension/domain • Classical psychometric criteria for the selection of items • Expert opinion for the reduction of dimensions/domains, items and response levels	Pulmonary hypertension	CAMPHOR	McKenna et al., 2008
	Overactive bladder	OAB-q (OAB-5D)	Yang et al., 2009; Young et al., 2009
	Asthma	AQLQ (AQL-5D)	Yang et al., 2011; Young et al., 2011
	Cancer	EORTC QLQ-C30 (EORTC-8D)	Rowen et al., 2011
		QLQ-C30 (QLQ-PBM)	Versteegh et al., 2012
	Epilepsy	NEWQOL (NEWQOL-6D)	Mulhern et al., 2012a
	Dementia	DEMQOL, DEMQOL-Proxy (DEMQOL-U, DEMQOL-Proxy U)	Mulhern et al., 2012b; Rowen et al., 2012)

ii. Development of health state classifications that are unidimensional or have unidimensional components [method developed for this thesis]			
Primary methodology used	**Condition**	**Original measure (new measure)**	**Relevant references**
• ± PCA for exploration of the dimensionality of the original measure • Rasch analysis and classical psychometric criteria for the construction of unidimensional health state descriptions and the selection of items and response levels • Expert opinion for the reduction of dimensions/domains and items	Common mental health problems	CORE-OM (CORE-6D)	Mavranezouli et al., 2011
	Diabetes	ADDQoL (DUI)	Sundaram et al., 2009 & 2010
	Flushing	FSQ	Young et al., 2010
	Arthritis	HAQ (HAQ-PBM)	Versteegh et al., 2012
	Multiple sclerosis	MSIS-29 (MSIS-PBM)	Versteegh et al., 2012
	Vision loss	NEI VFQ-25 (VFQ-UI)	Kowalski et al., 2012
B. Identification of plausible health state descriptions			
Primary methodology used	**Condition**	**Original measure (new measure)**	**Relevant references**
Cluster analysis for the identification of distinct patient severity groups	Depression	SF-12	Lenert et al., 1999 & 2000a; Sugar et al., 1998
	Schizophrenia	PANSS	Lenert et al., 2004; Mohr et al., 2004
C. Hybrid approach: development of health state classifications and identification of plausible health state descriptions			
Primary methodology used	**Condition**	**Original measure (new measure)**	**Relevant references**
• Rasch analysis for the selection/exclusion of items in each domain • Classical psychometric criteria for the selection of items • Expert opinion for the selection of the items • Cluster analysis for the identification of distinct patient severity groups	Rheumatoid arthritis	HAQ	McTaggart-Cowan et al., 2010

FA: Factor analysis; IRT: Item Response Theory; PCA: Principal Components Analysis

Expert judgement is required at each stage to interpret the results of all analyses undertaken and finalise decisions on the selection of the most appropriate dimensions, items and response levels.

Stage 1. Assessment of the dimensionality of the original measure

Factor analysis and principal components analysis

Assessment of the dimensionality of a measure can be achieved by factor analysis (FA) or principal components analysis (PCA). These are statistical techniques that can assess whether variables (e.g. items of a measure) form coherent subsets that are relatively independent from each other. Variables that are correlated with one another but at the same time are largely independent from other subsets of variables are combined into factors or components, respectively (Tabachnick & Fidell, 1996). The difference between FA and PCA is in the variance that is analysed. FA assumes that the variance in the measured variables can be decomposed into that accounted for by common factors, and that accounted for by unique factors. Subsequently, FA analyses only shared variance (covariance), accounted for by the common factors. PCA on the other hand analyses all the variance in the observed variables, both common and unique. PCA is a unique mathematical solution whereas most forms of FA are not unique (DeCoster, 1998; Tabachnick & Fidell, 1996).

FA and PCA can identify the number of reliable and interpretable factors (components) underlying the variables in a dataset. They can also estimate the extent of variance in a dataset that is accounted for by these factors and indicate the factors that account for the most variance within the dataset. A factor is more easily interpreted when several observed variables correlate highly with it and those variables do not correlate with other factors. Based on their properties, FA and PCA can be used to assess the dimensional structure of an instrument, to explore potential correlations between dimensions, and to suggest appropriate reductions in dimensions (Chatfield & Collins, 1980). The steps in both processes include selecting and measuring a set of items (variables) forming an instrument, preparing the correlation matrix between each pair of items, extracting a set of factors from the correlation matrix,

determining the number of factors, rotating the factors to increase interpretability, and interpreting the results (Tabachnick & Fidell, 1996). Interpretation of results requires personal judgement, because it is possible that the methods assign items belonging to the same conceptual dimension into different components, based on their level of 'difficulty'; this may occur because 'easy' variables (for example items capturing milder levels of disease) and 'difficult' variables (for example items with an ability to identify severe levels of disease) have higher correlations amongst themselves (Bond, 1994). Likewise, items belonging to different dimensions may be assigned to the same component if they are phrased in a similar way that is distinct to phrasing of other items (for example negatively versus positively worded items).

There are two major types of FA: exploratory and confirmatory. Exploratory FA aims to describe and summarise data by grouping together variables that are correlated. It examines whether there is an underlying pattern of scales amongst a set of questionnaire items. It is a tool for consolidating variables and for generating hypotheses about underlying processes. Confirmatory FA is a more sophisticated technique used in advanced stages of the research process to test a theory about latent processes (Tabachnick & Fidell, 1996). More specifically, confirmatory FA tests whether a specified set of constructs is influencing responses in a predicted way (DeCoster, 1998).

The number of factors (components) in a dataset can be estimated using various criteria (DeCoster, 1998; Tabachnick & Fidell, 1996). The Kaiser criterion relies on examination of the sizes of the eigenvalues in the correlation matrix; eigenvalues express the amount of variance in the data that is reproduced by a given factor. The number of significant factors equals the number of the eigenvalues that are above 1 (Kaiser, 1960). Another criterion is the scree test of eigenvalues plotted against factors (Cattell, 1966). Factors, in descending order, are arranged along the abscissa with eigenvalues as the ordinate. Usually the eigenvalue is highest for the first factor and moderate but decreasing for the next few factors before reaching small values for the last several factors. The last important factor lies at the point where the line drawn through the points changes slope. The scree test involves judgement as to

where discontinuity in eigenvalues occurs, especially when the sample size is small, communalities are low, and each factor has few variables with not particularly high loadings. A final method for identifying the number of significant factors is based on Horn's parallel analysis (Horn, 1965). This procedure involves generation of random datasets of uncorrelated variables that have the same number of cases and variables with the actual dataset. Subsequently, eigenvalues are computed for the correlation matrices of the original data and of each of the random datasets. Components whose eigenvalues estimated from the original data are greater than eigenvalues estimated from the random data should be retained. Horn's parallel analysis has been identified as the most accurate method for estimating the number of significant factors (Zwick & Velicer, 1986).

Rotation of factors / components is a process by which the solution is made more interpretable without changing its underlying mathematical properties. There are two general classes of rotation. Orthogonal rotation assumes that all factors are uncorrelated with each other; in this case, a loading matrix is produced. This is a matrix of correlations between observed variables and factors. The sizes of the loadings reflect the extent of the relationship between each observed variable and each factor. Oblique rotation assumes that there is correlation across the factors. In this case, several additional matrices are produced: the factor correlation matrix provides the correlations among the factors; the structure matrix presents the correlations between factors and variables; and the pattern matrix shows the unique relationships between each factor and each observed variable, uncontaminated by overlap among factors. In oblique rotation the meaning of factors is ascertained from the pattern matrix (Tabachnick & Fidell, 1996).

PCA played an important role in the identification and selection of dimensions to be retained in the first derivation of a health state classification from an existing measure, that is, the derivation of the SF-6D from the SF-36 health survey (Brazier et al., 1998 & 2002). The SF-36 consists of 36 items that belong to 8 different dimensions; the 36 items have different levels of response that are not comparable across the items (Ware et al., 1993). Brazier and

colleagues (2002) derived the SF-6D from SF-36 by selecting appropriate dimensions, items and levels from the initial instrument. This process was based on the results of a PCA that had been previously undertaken at the development of SF-12, a shorter form of SF-36 (Ware et al., 1995), examination of the correlations between the SF-36 items and between each of the SF-36 items and the whole measure, and expert judgement. The derived SF-6D classification consists of 6 single-statement dimensions (physical functioning, role limitations, social functioning, pain, mental health and vitality), covering 11 of the SF-36 items (some items of SF-36 were combined into a single item in SF-6D), with each statement having between 4 and 6 levels of response, determined by expert judgement. The SF-6D can describe 18,000 different health states, 249 of which were selected for the valuation survey using orthogonal arrays (Brazier et al., 2002). Similar work was undertaken to derive a health state classification from the SF-12 (Brazier & Roberts, 2004).

Kok and colleagues (2002) derived a health state classification from the International Prostate Symptom Score (IPSS) using PCA as the primary tool. IPSS consists of 7 questions about symptoms and one question assessing the impact of symptoms on patients' HRQoL. Each of the 8 questions has 6 levels of response. PCA undertaken on the 7 symptom items revealed that these belonged to 2 components that could be interpreted as 'obstructive symptoms' and 'irritative symptoms'. The 6 response categories of each component were merged into 3 levels by expert judgement. Effectively, the authors constructed a new 2-component measure consisting of two items (one capturing obstructive and the other capturing irritative symptoms) with 3 levels of response each that was possible to describe $3^2 = 9$ distinct health states, all of which were included in a valuation survey.

Kind and Macran (2005) and Lamers and colleagues (2007) derived a health state classification system from the Functional Assessment of Cancer Therapy-Lung (FACT-L) using the results of FA and a qualitative review of the items. FA was carried out to determine the dimensional structure of the measure and to identify the most representative items within each dimension. The qualitative review of items aimed at determining each item's importance

and suitability for use in a PBM. This process resulted in a health state classification with 6 dimensions (physical, emotional, functional, social/family, general symptoms and specific symptoms). Four of the dimensions (physical, emotional, functional and specific symptoms) contained 2 items each, so that the new measure contained 10 items with 2 response levels (yes/no) each. The health state descriptions that were used in valuation contained only one item per dimension (i.e. 6 items in total), so that 2 different versions of the health state classification were developed. In total, the system described 2^6 = 64 distinct health states. Two subsets of 10 health states using the 2 versions of the health state classification were selected, using orthogonal arrays, for 2 valuation surveys that were conducted in the UK (Kind & Macran, 2005) and Denmark (Lamers et al., 2007).

FA and PCA have comprised the first step in the process of deriving several other health state classifications from existing measures, as described in the sections that follow.

Stage 2. Selection of items and item response levels for inclusion in the health state classification

Selection of items of a measure for inclusion in a health state classification can be made using classical psychometric criteria. Item response theory (IRT) has also been reported as a tool in the selection of items. Reduction in item response levels can be made based on expert judgement. More recently, selection of appropriate items and item response levels has been achieved using Rasch analysis, a mathematical model that belongs in the family of the IRT models.

Classical psychometric criteria

As discussed in Chapter 1 (section 1.2.4), classical psychometric criteria are widely used for the assessment of outcome measures. These include the appropriateness of a measure, its reliability (relating to the measure's internal consistency and reproducibility), validity (consisting of face validity, content validity, construct validity, criterion validity and predictive validity), responsiveness, precision, interpretability, acceptability and feasibility. In

addition to the evaluation of whole measures, a number of these criteria have also been used for the selection of 'best-performing' items for the derivation of a scale from a larger questionnaire. The most commonly used classical psychometric tests that have been used for the assessment of individual items considered for inclusion in a new measure derived from an existing longer scale are the following:

- internal consistency, expressed by the correlation of an item with the total scale score, or the total score of the dimension it belongs to
- construct validity, as assessed by known groups validity (the item's ability to distinguish between groups with different levels of the severity of the condition) and convergent validity (the item's correlations with other variables that have been designed to measure the same construct)
- responsiveness over time, usually assessed from the item's ES (the item's change score divided by the standard deviation of the score at baseline), or from the item's SRM (the item's change score divided by the standard deviation of the change score)
- reproducibility, as assessed by the item's test-retest reliability
- precision of an item, reflected in the distribution of responses across its response levels, which can be assessed by the magnitude of ceiling or floor effects
- acceptability of an item to respondents, reflected in the rate of missing data.

A number of health state classifications have been successfully derived from longer measures using primarily classical psychometric criteria supplemented by expert judgment, including a menopause-specific health quality of life questionnaire (Brazier et al., 2005b), the King's Health Questionnaire (KHQ) health state classification for urinary incontinence (Brazier et al., 2008), and a health state classification derived from the Sexual Quality of Life questionnaire (SQOL) (Ratcliffe et al., 2009). In all cases, a range of classical psychometric criteria among those described above were used in combination with expert judgement in order to initially exclude inappropriate items (for example items that lacked face validity or items with relatively poor performance) and

subsequently select best-performing items from the existing measures, so as to develop concise health state classifications. In the case of the KHQ, expert judgement was used prior to this process, to exclude dimensions that were not appropriate for or directly relevant to a HRQoL measure.

The SQOL health state classification (SQOL-3D) consisted of 3 items, each with 4 response levels, corresponding to the original 3 dimensions of the SQOL (sexual performance, sexual relationship and sexual anxiety), and therefore it described 4^3=64 possible health states (Ratcliffe et al., 2009). The menopause health state classification comprised 7 single-item dimensions with 3 or 5 response levels each, defining 6,075 potential health states (Brazier et al., 2005b). The KHQ health state classification consisted of 5 single-item dimensions with 4 response levels each, describing 4^5=1,024 health states (Brazier et al., 2008). Given the large number of potential health states described by the menopause and the KHQ health state classifications, a number of the health states were selected for the valuation survey using orthogonal arrays.

Item response theory

IRT has been reported as a tool in the selection of items from a large questionnaire for inclusion in a health state classification. IRT comprises a family of mathematical models that are useful in the design and analysis of psychological and educational measures (Weiss & Yoes, 1991). IRT models assess how much of an attribute a person possesses, based on the person's responses to items of a scale designed to measure the attribute (Baker, 2001). IRT models are designed to predict the probability of affirming an item, depending on the person's amount of the attribute and a number of item parameters (Harvey & Hammer, 1999; Streiner & Norman, 1995). The simplest IRT model is the one-parameter logistic model (Rasch model), which assumes that only a single item parameter is required to predict a person's response to an item. This is the 'difficulty' of an item (that is, the amount of the attribute the item is able to capture). The 2-parameter logistic model considers a second item parameter, that of the 'discriminative ability' of an item, reflecting the fact that some items on a scale have stronger or weaker relations than others to the attribute being assessed. Finally, the 3-parameter logistic model takes into

account that persons with a very low amount of the attribute may still affirm an item due to pure chance or due to a 'social desirability' for a positive response to the item.

Stevens and colleagues (2005) used IRT in addition to classical psychometric tests in order to derive a health state classification for children with atopic dermatitis from a larger questionnaire. IRT was used to select items that represented different severities of impact on the child. The resulting health state classification consisted of 4 items with 2 response levels each, and therefore it formed 2^4=16 potential health states, all of which were included in the valuation survey.

Rasch analysis

The Rasch model is by large the most commonly used IRT model for the derivation of health state classifications from existing measures. Rasch analysis has been used in combination with traditional psychometric criteria for selection of items and item response levels. Rasch analysis is a statistical measurement approach for examining the relationship between people's attributes, such as knowledge, quality of life or morbidity, and ordinal scales designed to measure such attributes. It is based on the principles of the Rasch model (Rasch, 1960) according to which the outcome of an encounter between a person and an item is exclusively governed by the product of the person's 'ability' (i.e. the person's 'amount' of the attribute) and the item's 'difficulty' (i.e. how much 'quantity' of the attribute the item is able to capture) (Tennant & Conaghan, 2007). The model is a probabilistic form of Guttman scaling, a deterministic pattern that expects a strict hierarchical ordering of items (e.g. from low to high difficulty) such that if a person has affirmed an item of a given level of difficulty, then all easier items on the scale should also be affirmed (Guttman, 1950). The Rasch model relaxes this proposition by stating that if a more difficult item is affirmed, then there is a high probability that easier items will also be affirmed (Tennant & Conaghan, 2007).

The Rasch model assumes that the probability of a given person affirming an item is a logistic function of the relative distance between the item's location (determined by the item's difficulty) and the person's location (determined by

the person's ability) on a continuous scale with interval properties (Pallant & Tennant, 2007). For dichotomous data, this can be mathematically expressed as:

$$p_{n_i} = \frac{e^{(\theta n - b_i)}}{1 + e^{(\theta n - b_i)}}$$

where p_{ni} is the probability that person n will affirm item i, θ is the person's ability, and b is the item's difficulty. Thus, the probability of a 'correct' (affirmed) response increases as the ability of a person increases, and the difficulty of an item decreases (theory of conjoint measurement) (Bond & Fox, 2007).

Rasch analysis can convert ordinal scale scores into measurements of the attribute on a continuous (latent) scale with interval properties, with the logit (log odds unit) as the unit of measurement (Bond & Fox, 2007; Tennant & Conaghan, 2007; Tennant et al., 2004). The logit is the distance along the line of the scale that increases the odds of a person affirming an item of average difficulty by a factor of 2.718. The Rasch model demonstrates what the expected responses to items should be (according to each person's ability and each item's difficulty), if interval scale measurement is to be achieved (Tennant & Conaghan, 2007). Subsequently, the observed responses are compared with the expected ones in order to assess whether the differences between observed and expected scores ('residuals') are significant and whether the examined dataset (in terms of both persons *and* items) fits the Rasch model (Tesio, 2003). When a dataset fits the Rasch model, then Rasch analysis allows prediction of a person's responses to each item based exclusively on the person's ability and each item's difficulty (Tennant et al., 2004).

Rasch analysis assigns individual persons and items on different points (or 'locations') along the Rasch model logit scale, according to each person's ability (reflected in the percentage of items affirmed by the person) and each item's difficulty (reflected in the percentage of persons affirming the item). Assignment of persons and items across the scale presupposes that the ability of a person and the difficulty of an item are independent from each other

('separability theorem') (Bond & Fox, 2007). Each location along the continuous scale corresponds to a 'Rasch model logit value', with higher values expressing more difficult items and more 'able' persons (i.e. persons with higher amounts of the attribute). Respondents with the same ability on an attribute (and therefore the same total score on the ordinal scale) are assigned the same Rasch model logit value. The Rasch model logit scale is centred on zero logit; the latter represents the item of average difficulty in the scale (Tennant et al., 2004). Assignment of persons to different points along the scale leads to generation of groups of respondents with different levels of ability in the measured attribute (Bond & Fox, 2007).

The Rasch model is characterised by unidimensionality and local independence of items. Unidimensionality means that all items of a scale fitting the Rasch model capture a single attribute. Local independence of items means that, once the 'Rasch factor' (i.e. the attribute) has been removed, there should be no further associations (other than random associations) between the items of the scale (Tennant & Conaghan, 2007; Tennant et al., 2004). Local dependence may arise when the scale is multidimensional (and therefore there are correlations between items beyond those that are attributable to the Rasch factor) or when there is response dependency between some items (i.e. when there is a logical relationship between the items so that the response to one item determines the response to another item).

Although originally Rasch analysis was developed for application in dichotomous items, the theory has been extended for the analysis of polytomous categorical items (Andrich, 1978). The rating scale Rasch model is used when the polytomous items have the same response levels whereas the partial credit Rasch model is used when polytomous items have different response levels. Although these two models differ in the parameterisation and the number of degrees of freedom, they do not differ in the structure and the response process for a person responding to an item (Luo, 2005).

Rasch analysis can be used to assess the following characteristics of an instrument and its items (Pallant & Tennant, 2007; Tennant & Conaghan, 2007):

- Overall goodness of fit in the Rasch model: Rasch analysis assesses the degree of the discrepancy between observed and expected responses. Item-person interaction statistics are expressed by a Z score representing a Z-standardised normal distribution. If items and persons fit the model, then the mean of the distribution is expected to approximate zero and the standard deviation to reach one. The item-trait interaction measures, by the means of chi-squared statistics, whether data fit the Rasch model for discrete groups of responders that represent different levels of ability (class intervals) across the attribute. A significant chi square indicates that the hierarchical ordering of the items varies across the attribute, thus compromising the required property of invariance.

- Individual item and person fit: relevant chi-squared statistics demonstrate whether distinct items and persons fit the Rasch model. Item and person fit residuals examine the amount of variability between the expected and observed responses for each item and each person separately.

- Threshold ordering of polytomous items: thresholds are the points (locations) on the latent scale where the probability of response in adjacent response levels is equally likely (50%). The Rasch model expects thresholds to increase with increasing difficulty of adjacent response levels (i.e. the threshold between adjacent response levels 2 and 3 should be further on the scale from the threshold between adjacent response levels 1 and 2), so that the probability of obtaining a higher item score increases as the ability of a respondent increases; this indicates that respondents are able to distinguish between adjacent response levels (ordered thresholds). Disordered thresholds are observed when an item score is likely to decrease as respondent's ability increases; this means that respondents cannot distinguish between adjacent levels of response of this item. In order to obtain items with ordered thresholds, adjacent response levels of

items with disordered thresholds should be collapsed (merged) and checked for threshold ordering in a subsequent Rasch analysis.

- Differential Item Functioning (DIF): this occurs when different sub-groups within a study sample (discriminated by age, gender, or other socio-demographic characteristics) behave differently and give persistently different responses to one or more items, despite of having equal levels of the attribute being measured. Uniform DIF exists when the sub-groups show a consistent systematic difference in their responses to an item, across the whole range of the attribute being measured. Non-uniform DIF occurs when there is non-uniformity in the differences across sub-groups (i.e. patterns of difference vary across different locations of the scale). DIF can be a cause of misfit to the Rasch model. While non-uniform DIF cannot be dealt with, uniform DIF can be resolved by splitting the item demonstrating DIF and creating unique 'sub-items' corresponding to each sub-group with different baseline characteristics for which DIF was identified (Brodersen et al., 2007).

- Targeting of persons and items: Rasch analysis can assess whether an instrument can capture the whole range of symptom severity observed in the study population. In a well-targeted instrument, the average location of the study population should coincide with the average location of items; in addition, no floor or ceiling effects should be observed.

- Reliability: this is expressed by the person separation index (PSI), which measures the discriminative ability of the instrument across different groups of responders and is equivalent to Cronbach's α in traditional test theory (Cronbach, 1951).

- Unidimensionality and local independence of items: these can be tested by a variety of methods including independent *t*-tests (Smith, 2002) and PCA of the fit residuals (Wright, 1996).

A range of these criteria have been used in combination with standard psychometric tests for the selection of items and the reduction of response levels in a number of studies that derived health state classifications from existing CSMs. McKenna and colleagues (2008) were the first to employ Rasch analysis for this purpose. The authors reduced the 25-item Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) into a health state classification that consisted of 6 items, each with 2 or 3 response levels, belonging to 4 domains. Selection of items was based on the following criteria:

- the item's loading onto its domain as revealed by FA
- the logit location of each item in Rasch analysis that was conducted separately on each domain of the CAMPHOR (items with extreme location were candidates for exclusion)
- the percentage affirmation of each item (items affirmed by a very small or very large percentage of respondents were excluded)
- the correlation of each item with a general health perception variable that was predicted by the CAMPHOR responses by ordinal regression (items with high correlation were candidates for inclusion)
- expert opinion, which was used to assess the face validity of items and the coverage of relevant aspects of HRQoL in the new health state classification

The methodology first described by McKenna and colleagues (2008) was refined and standardised in a number of studies that used primarily Rasch analysis supplemented by standard psychometric criteria to derive multidimensional health state classifications from existing CSMs, including the derivation of the OAB-5D from the Overactive Bladder Questionnaire (OAB-q) (Yang et al., 2009; Young et al., 2009), the AQL-5D from the Asthma Quality of Life Questionnaire (AQLQ) (Yang et al., 2011; Young et al., 2011), the EORTC-8D from the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30) [in two separate studies by Rowen and colleagues (2011) and Versteegh and colleagues (2012)], the NEWQOL-6D from a quality of life measure for epilepsy (NEWQOL) (Mulhern et al., 2012a), and the DEMQOL-U and DEMQOL-Proxy-

U from 2 quality of life measures for dementia rated by patients (DEMQOL) and carers (DEMQOL-Proxy), respectively (Mulhern et al., 2012b; Rowen et al., 2012), All these studies used broadly the same methodology that can be summarised as follows (Young et al., 2009):

Step I: PCA was used to establish the domain structure of the original measure; eigenvalues, scree plots and the rotated component matrix were examined for this purpose. PCA was also used to identify potential correlations between items and domains, which informed the choice of domains for inclusion in the health state classification. Moreover, items that loaded either on no component or on more than one components were removed from analysis.

Step II: Rasch analysis was undertaken separately in each domain to reduce the number of items by excluding unsuitable items. The following criteria were used:

- Threshold ordering: where items had disorder thresholds, ordering was achieved by merging adjacent item levels using an item-by-item approach; these items were not considered further for inclusion in the health state classification, as it was deemed that they no longer captured the full-range severity of the original measure; nevertheless, they were not removed from subsequent Rasch analyses as threshold ordering might result in the item's fitting in the Rasch model.
- DIF: items demonstrating DIF were of limited value for making cross-population comparisons and therefore were not considered for inclusion in the health state classification; however, these items were separated into different person factors (for which DIF was observed) and were retained in subsequent Rasch analyses.
- Goodness of fit: the overall model fit was assessed by the item-trait interaction statistics and the PSI; individual item fit statistics and fit residuals were also examined. The item with the poorest individual fit statistics was removed from the model and Rasch analysis was re-run. This process was repeated until all non-fitting items were removed and the

overall item-trait interaction showed a satisfactory model fit. All non-fitting items were not considered for inclusion in the health state classification.

Step III: The results of the Rasch analysis undertaken in the previous step and standard psychometric criteria were used to select one best-fitting item per domain, aiming to select items capturing the full range of condition severity. The following criteria were used:

- Rasch analysis criteria:
    - the spread of item response levels across the logit scale, as depicted in threshold probability curves which show the distribution of item response levels across the logit scale (the wider the spread, the easier for respondents to distinguish between adjacent response levels)
    - individual item fit statistics
- Classical psychometric criteria:
    - acceptability (rate of missing data)
    - internal consistency (correlation between item and domain scores)
    - distribution of responses (e.g. magnitude of ceiling or floor effects)
    - responsiveness to change over time (assessed using SRM)

Step IV: Following selection of one item per domain for inclusion in the health state classification, the results of Rasch analysis undertaken in step II were used in order to reduce the response levels of each item, so that health state descriptions were more concise and easier to process in a valuation survey. This was achieved by inspecting the threshold probability curves: response levels with close thresholds were candidates for response level collapsing. Moreover, response levels with low percentage of responses were candidates for merging with an adjacent response level.

Step V: Validation of all the previous steps in a separate dataset.

In all steps expert opinion was used to assess the results of the statistical analyses, evaluate qualitative features of the items and make final decisions regarding the inclusion of items in the final health state classification.

The 5-step methodology described above was first adopted by Young and colleague (2009), for the derivation of the OAB-5D from OAB-q, a measure specific to overactive bladder. OAB-5D consists of 5 single-item dimensions with 5 response levels each that were established by PCA (urge, urine loss, sleep, coping and concern). The OAB-5D can describe 5^5 = 3,125 potential health states; of these, 98 were selected for valuation using a balanced statistical design (Yang et al., 2009).

Similarly, Young and colleagues (2011) derived the AQL-5D from AQLQ. AQLQ consists of 32 items with 7 response levels each, which form 4 distinct domains (symptoms, activity limitations, emotional function and environmental stimuli). Extra items not included in AQLQ and corresponding to a fifth domain (sleep) were considered for inclusion in AQL-5D because sleep was deemed to be an important aspect of HRQoL in patients with asthma. AQL-5D consists of 5 single-item domains, each with 5 response levels. The system is possible to describe 5^5 = 3,125 potential health states of which 99 were selected for valuation using a balanced statistical design (Yang et al., 2011).

The same principles were adopted at the derivation of EORTC-8D from EORTC QLQ-C30 (Rowen et al., 2011). EORT QLQ-30 is a cancer-specific measure that contains 30 questions covering the most common cancer symptoms (such as pain, fatigue, nausea and vomiting) and various aspects of functioning (including physical, role, social, emotional and cognitive). The QLQ-C30 is summarised using 14 2-item scales, each representing a particular symptom or aspect of functioning, plus one 2-item global quality of life scale. After initial exclusion of 3 items that related to global quality of life and financial impact due to their inappropriateness for inclusion in a PBM, the remaining 27 items were subject to the process described in Steps I-V. The resulting EORT-8D comprised an 8-dimensional health state classification (capturing physical functioning, role functioning, pain, emotional functioning, social functioning, fatigue and sleep disturbance, constipation and diarrhoea, and nausea) with each dimension being represented by one item; all items have 4 response levels, except the physical functioning item which has 5

response levels. The new measure can describe 81,920 health states, 85 of which were selected for valuation using an orthogonal statistical design.

Mulhern and colleagues (2012b) used the process described in Steps I-V to derive patient-reported and carer-reported health state classifications for dementia from DEMQOL and DEMQOL-Proxy, respectively. DEMQOL consists of 28 items with 4 response levels each, tapping health and well-being, cognitive functioning, social relationships, daily activities and self-concept. DEMQOL-Proxy, which was designed to enable measurement of HRQoL in people with severe dementia, shares the same conceptual framework with DEMQOL and contains 31 items with 4 response levels each. Both measures include a global quality of life item that does not contribute to the overall measure score. In step I of the process, PCA identified 5 dimensions in DEMQOL (positive emotion, memory, relationships, negative emotion and loneliness) and 4 dimensions in DEMQOL-Proxy (all the above except loneliness). Following application of Rasch analysis and psychometric criteria, 2 health state classifications were derived, respectively: the 5-dimensional DEMQOL-U with 4 response levels in each single-item dimension that generate $4^5 = 1{,}024$ possible health states and the 4-dimensional DEMQOL-Proxy-U with 4 response levels in each single-item dimension that describe $4^4 = 256$ potential health states. A representative sample of states was selected from each health state classification for valuation using a block design, which led to the generation of combinations of states; each respondent in the survey valued a block of seven mixed states plus the worst state (Rowen et al., 2012).

Mulhern and colleagues (2012a) described the same 5-step process at the derivation of NEWQOL-6D from NEWQOL. NEWQOL includes a range of measures validated for general use across a range of conditions and specifically for epilepsy; of these, a subset of 82 items was considered for the derivation of the new health state classification. NEWQOL-6D has 6 single-item dimensions with 4 response levels each, describing $4^6 = 4{,}096$ distinct health states. A sample of 50 health states was selected using an orthogonal array for consideration in a valuation survey.

Finally, Versteegh and colleagues (2012) derived a health state classification from the QLQ-C30 using PCA, Rasch analysis, classical psychometric criteria and expert opinion. The resulting health state classification comprised 8 items belonging to 5 dimensions (physical functioning, vitality, mental functioning, discomfort, pain). Selection of health states for the valuation survey was based on a level-balanced design, meaning that all response levels of each item were seen with the same frequency within the selected health states; the latter covered the entire spectrum of severity of symptoms.

Stage 3. Validation of the new health state classification

Validation of a newly developed health state classification is an essential step that provides confirmation that the final set of dimensions, items and response levels constitutes an optimal solution. This can be achieved by repeating the process used for the derivation of the health state classification on a separate study sample and/or on data from the same study sample collected at a different time point. This method has been employed at the validation of OAB-q (Young et al., 2009), AQL-5D (Young et al., 2011), EORTC-8D (Rowen et al., 2011) and NEWQOL-6D (Mulhern et al., 2012a).

Discussion – strengths and limitations of the health state classification approach

The health state classification approach comprises a useful method for deriving health state descriptions from existing longer measures usually consisting of multiple dimensions, items and response levels. Application of statistical methods such as PCA, Rasch analysis and standard psychometric criteria combined with expert judgement allows identification of the most appropriate dimensions, items and response levels, so that the health state classification is able to capture a variety of HRQoL aspects that are relevant to the study population, across a range of HRQoL levels.

Health state classifications have been routinely derived from measures consisting of multiple dimensions or domains with little or no correlation between them. Ideally, health state classifications should retain this

multidimensional structure and include items that behave independently. Apart from the desired ability of health state classifications to tap as many relevant HRQoL aspects as possible, 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 described by a health state classification. As described in Chapter 1 (section 1.3.1), the major approach for generating and selecting health states from a health state classification for use in a valuation survey relies on the use of conventional statistical approaches such as orthogonal arrays and balanced designs; such techniques have been used in order to select health states for valuation from the multidimensional EQ-5D (Dolan et al., 1996) and SF-6D (Brazier et al., 2002). Alternatively, valuation of HUI-3 was based on MAUT, which involved valuation of 'corner' states, where one dimension is at the worst level and all others are at the best level (Feeny et al., 2002). All these techniques employed for the generation of health states treat items independently, as separate statements. If there is indeed no correlation between the items of a health state classification, then any combination of items will result in the description of a plausible health state.

A major problem arises when items in a health state classification tap the same or highly correlated dimensions and therefore cannot be treated independently when generating health states. In such cases, some of the health states generated using standard approaches 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 when the original measure has narrow scope and is characterised by high correlations between its dimensions and items. Thus, derivation of health state classifications using the methodology described earlier and subsequent use of standard approaches for generating random health states (e.g. using orthogonal or block designs) or 'corner' states for valuation may not be appropriate if the original measure is largely unidimensional or consists of highly correlated dimensions and items.

An alternative approach is therefore required for the derivation of health state classifications from measures that are unidimensional or are characterised by high correlations between their dimensions / domains and items.

3.3.2 Derivation of health state descriptions in the form of vignettes – the clustering-based approach

A different approach for deriving health state descriptions from existing measures is to construct health states from item responses that are observable in the study population by grouping (clustering) patients according to their level of symptom severity.

Use of cluster analysis for the identification of distinct patient severity groups that lead to the construction of plausible health states

The basic aim of cluster analysis is to reveal natural groupings (or clusters) within a set of individuals (Chatfield & Collins, 1980). This is achieved by allocating a set of individuals to a set of mutually exclusive groups based on selected characteristics, such that individuals within a group are similar to one another while individuals in different groups are dissimilar. An application of cluster analysis is the grouping of patients in a dataset according to their severity of symptoms. The resulting clusters are groups of patients with different levels of symptom severity, which indicate distinct health states of a condition / disease area.

This approach was first described by Sugar and colleagues (1998) who conducted k-means cluster analysis using the mental and physical health composite scores of SF-12 obtained from patients with depression, in order to assign them into groups of different symptom severity. K-means cluster analysis attempts to identify relatively homogeneous groups of cases based on their characteristics, using an algorithm that can handle large numbers of cases. The algorithm requires the researcher to specify the number of clusters; once the number is selected, the algorithm specifies cluster membership. The process resulted in specification of 6 distinct patient groups corresponding to 6 respective health states covering 2 dimensions, i.e. mental and physical health (the 6 health states were ‘near normal health’, ‘mild mental and physical impairment’, ‘severe physical impairment’, ‘severe mental impairment’, ‘severe mental and moderate physical impairment’ and ‘severe mental and physical impairment’). Subsequently, the authors examined the distribution of patients’ responses to SF-12 in each cluster and found that, for any item, one or two

levels of response accounted for at least 50% of patient responses in a cluster. By combining these 'popular' item responses, the authors developed 6 health state descriptions for depression that were clinically meaningful; these health state descriptions formed vignettes that were later valued by patients with depression (Lenert et al., 1999 & 2000a).

A similar exercise was carried out to construct health states for schizophrenia using clinicians' ratings on the PANSS (Mohr et al., 2004). Previously conducted FA had identified 5 domains within the scale (positive symptoms, negative symptoms, cognitive impairment, mood disorder and hostility/aggression). K-means cluster analysis was then conducted on the sum of standardised PANSS scores within each domain to identify clusters of patients with similar profiles of schizophrenic symptoms; results of cluster analysis were compared with a conceptual framework of health states developed by an expert panel. Final health states were determined by combining profiles of schizophrenic symptoms from all PANSS domains, after assessing the empirical results in conjunction with the conceptual framework. This process resulted in the formation of 8 plausible health states with varying levels of positive, negative and cognitive impairment, ranging from mild to extremely severe symptoms, that are observable in the study population. These health states were subsequently valued by a sample of the general population in the US (Lenert et al., 2004).

Hybrid approach: derivation of a health state classification followed by cluster analysis for the construction of health states within the classification

McTaggart-Cowan and colleagues (2010) used a hybrid approach in order to derive distinct health states from the Health Assessment Questionnaire (HAQ), a measure for patients with rheumatoid arthritis. HAQ is an instrument with a high number of components and items. The authors selected a 20-item component of the instrument that was appropriate to form the basis of a PBM; these 20 items have 4 response levels each and cover 8 domains relating to patient's ability to complete daily tasks, such as dressing and grooming, arising, eating, walking, personal hygiene, reach, grip, and other activities.

Before any analysis was conducted, it was proposed that the new health state classification include 5 items, so as to be handled without difficulty by respondents in a valuation survey. The choice of items was made based on Rasch analysis and classical psychometric criteria following the approach proposed by Young and colleagues (2009) that was described earlier. The resulting instrument comprised 4 items with 4 response levels each. Subsequently, the authors aimed to produce a low number of plausible health states (3-4) that could cover a range of symptom severity in rheumatoid arthritis but at the same time could be easily managed by respondents at a valuation survey. For this reason k-means cluster analysis was conducted using the approach suggested by Sugar and colleagues (1998) to group respondents into distinct health states described by HAQ. Analysis indicated a solution of 3 health state clusters of varying severity of rheumatoid arthritis, ranging from very mild to severe. Further to this, the pain and discomfort dimension of EQ-5D was incorporated into the 3 health state clusters resulting in 3 health state descriptions (3 clusters) with 5 items each (4 HAQ items plus the pain and discomfort EQ-5D item). The 3 health state descriptions of rheumatoid arthritis were concise, plausible, and amenable to valuation.

Discussion – strengths and limitations of the clustering-based approach

The clustering-based approach for deriving health state descriptions from existing measures results in the construction of health states that are made up of frequent item responses that have been observed in the study population and are therefore clinically meaningful.

The main advantage of the clustering-based approach is that, in contrast to the health state classification approach, it does not require independence between the dimensions of a health state classification; the clustering-based approach allows construction of plausible health states and can therefore be employed for the development of PBMs from measures with few and highly correlated dimensions, where conventional approaches for generating health states are not appropriate. A limitation of the approach is that k-means cluster analysis uses arbitrary cut-off points for cluster identification and therefore it requires substantial input from experts. Another limitation of the approach as employed

in the studies described above was that clustering was based on patients' composite scores and not individual item responses. It is therefore possible that each cluster included patients with a wide range of individual item responses rather than a homogeneous patient group in terms of clinical presentation. Moreover, health descriptions were constructed by combining the most frequent scores/responses for every domain of the original measure in each cluster. However, these descriptions did not necessarily reflect the most frequent score / item response combinations in the study sample; what's more, it is possible that they did not form health states actually observed in the study population. A final drawback of the approach is that it results in a limited number of health states, thus potentially not covering all states that are routinely observed in the study population Nevertheless, this approach remains a strong alternative to the health state classification approach in situations where items of a questionnaire do not behave independently and therefore some of the potential health states derived from combinations of item statements are not plausible.

3.4 Conclusion

This chapter presented the results of a systematic review on the methods proposed in the literature for the derivation of health state descriptions that are amenable to valuation from existing measures. The aim of the review was to identify appropriate methods that can be used for the derivation of a health state descriptive system from the CORE-OM. The review revealed that there are two main approaches for this purpose:

The health state classification approach is the most widely reported in the literature. The approach uses a range of statistical techniques such as PCA, Rasch analysis and traditional psychometric tests to choose appropriate dimensions, items and response levels from the original measure for inclusion in a health state classification. The subsequent selection of health states for valuation by standard statistical designs presupposes that the health state classification is multidimensional with no correlations between its items. In principle, this methodology cannot be used to derive health state classifications from measures that are largely unidimensional or are characterised by considerable correlations between their components,

because it may result in the selection of health states that are implausible and contain contradictory statements.

The alternative clustering-based approach groups patients according to their symptom severity and uses these groupings to construct health state descriptions of varying severity. Its advantage is that it creates plausible health states by combining frequent responses of the patient population to the original measure. Therefore, it is appropriate to use in unidimensional measures or measures with high correlations between their items. However, the approach only identifies elements of health states that subsequently need to be put together to construct a full state and it cannot guarantee that the combinations of these elements, i.e. the resulting health states, are actually observed in the patient population. Moreover, this process results in a limited number of health states.

The decision on the approach to adopt for the derivation of a health state classification from the CORE-OM depends on the dimensionality of CORE-OM and the presence or absence of correlations between its items. As discussed in the next chapter, previous work has shown that CORE-OM does not have a clear multidimensional structure, suggesting that the clustering-based approach may be more appropriate to adopt. On the other hand, the review of the properties of the Rasch model suggests that it may be possible to derive a health state descriptive system by applying Rasch analysis on the entire CORE-OM if this is unidimensional or has a strong unidimensional component; following this process it is possible to generate plausible health states for the valuation survey by identifying groups of respondents that have been assigned to different points along the Rasch logit scale according to the severity of their symptoms. These considerations led to the decision to use the mainstream health state classification approach in order to derive health state descriptions amenable to valuation from the CORE-OM, by employing PCA, Rasch analysis and standard psychometric criteria.

Details on the methodology that was employed for the derivation of a health state classification from the CORE-OM are provided in Chapter 4.

Chapter 4. Methods used in this thesis for the derivation of a health state classification from the CORE-OM

4.1 Introduction

This chapter describes the methodology employed in this thesis in order to derive a health state classification that is amenable to valuation from the CORE-OM, which was the first stage in the development of a condition-specific PBM for people with common mental health problems. Analysis of CORE-OM data aimed at the selection of appropriate domains, items and response levels from the CORE-OM, so as to construct a concise and, at the same time, comprehensive health state descriptive system, able to capture a broad range of elements and levels of HRQoL in people with common mental health problems, and with minimum loss of information relative to the original 34-item measure.

The CORE-OM has been shown to comprise a largely unidimensional measure that is characterised by high correlation across its conceptual domains and items. Previously undertaken exploratory FA indicated that the 34 items load on 3 components, one including mainly the negatively worded items, one made up of the positively worded items, and one containing the risk items (Evans et al., 2002). Examination of the correlation across the instrument domains revealed 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 q 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 q value = 0.64 in a clinical sample; 0.44 in a non-clinical sample). These findings indicate that the mainstream methodology used to develop typically multidimensional health state classifications described in Chapter 3 may not be appropriate for the derivation of a new health state descriptive system from the CORE-OM; this is because a health state classification derived from the CORE-OM will also contain highly correlated items, thus entailing the danger that some of the

resulting health states generated using standard approaches may be implausible. Nonetheless, in order to derive a health state classification from the CORE-OM it was decided to use overall the same standard techniques that have been previously described in the literature for this purpose, but in such a way that the absence of clear multidimensionality of the CORE-OM is taken into account and that despite of the (unavoidable) incorporation of potentially highly correlated domains and items in the new measure, identification and selection of plausible health states is ensured.

Rasch analysis had a central role in this process, which effectively utilised the principle of unidimensionality underpinning the Rasch model (Tennant & Conaghan, 2007; Tennant et al., 2004) and its ability to generate plausible health states by identifying groups of respondents with distinct levels of symptom severity that have been assigned to different points along the Rasch logit scale. In summary, the process of deriving a new health state descriptive measure from the CORE-OM involved 4 steps, similar to those that have been described in the literature (Young et al., 2009):

1. Exploration of the dimensionality of the CORE-OM and correlations between its domains. The dimensionality of the CORE-OM was explored by undertaking PCA.

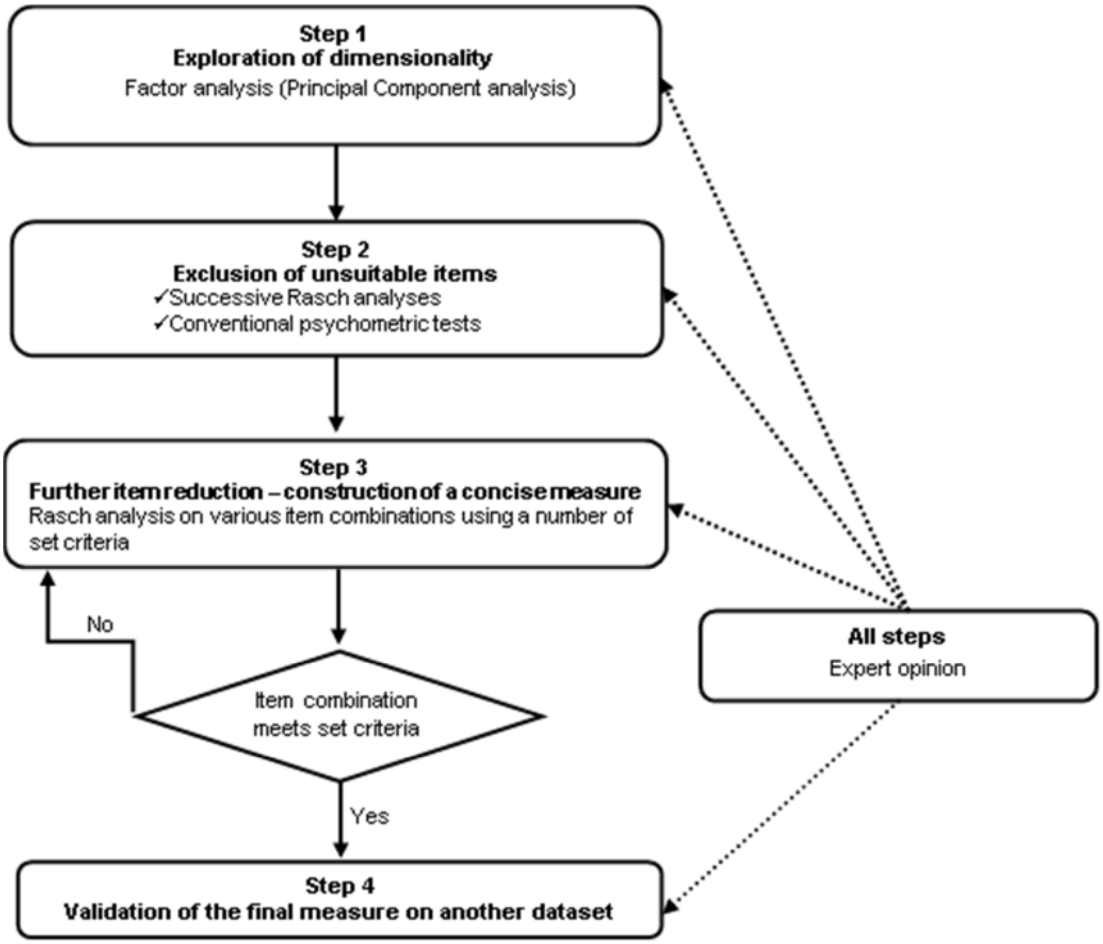
2. Investigation of the appropriateness and suitability of items and levels of response for inclusion in a health state classification, using predominantly a series of Rasch analyses and secondarily standard psychometric methods. Items that were deemed unsuitable according to a number of set criteria were excluded from further consideration.

3. Selection of items among those deemed suitable for inclusion in the final measure, using primarily Rasch analysis, leading to the development of a new health state classification.

4. Validation of the new health state descriptive system by repeating the above process on a different dataset.

In all steps of the process, an advisory group that was set up specifically for this purpose provided expert advice on the overall appropriateness of the methods employed, as well as on the interpretation and the clinical relevance of the findings. A flow diagram of this process is presented in Figure 6. Detailed description of the four steps of the process, an overview of the membership and the role of the advisory group and details on the dataset used in the analysis are provided in the remainder of the chapter.

Figure 6. Flow diagram of the process of deriving a new health state classification from the CORE-OM



4.2 Steps in the derivation of a health state classification from the CORE-OM

This section provides details of the 4-step approach that was used to derive a health state classification from the CORE-OM. The approach is similar to that described by Young and colleagues (2009) and adopted by many other studies with the same objective, as discussed in section 3.3.1 of Chapter 3.

4.2.1 Step 1. Exploration of the dimensionality of the CORE-OM

Given the indications of previous research for high correlations between the conceptual domains of the CORE-OM in clinical and non-clinical populations (Evans et al., 2002), this step aimed at further exploring the dimensionality of the CORE-OM and investigating potential strong correlations between its domains. Investigation of correlation structures would then determine whether the derivation of a health state classification system from the CORE-OM and subsequent generation of health states using a standard statistical design would be inappropriate, since this approach is likely to lead to generation of implausible health states when used in measures with highly correlated dimensions, as discussed in Chapter 3. Furthermore, this stage aimed to identify major domains within CORE-OM that should be ideally represented in the final instrument. Therefore, analysis undertaken at this stage intended to

- examine the correlation between underlying domains of CORE-OM and
- group the CORE-OM items into distinct domains so that, at next stages, the most suitable items from each domain were candidates for inclusion in the new health state classification.

The dimensionality of the CORE-OM was explored by undertaking PCA. PCA has been suggested as a tool in the development of new polytomous scales in order to provide early indications of dimensionality before Rasch analysis is attempted (Tennant & Pallant, 2006).

Principal components analysis

PCA was undertaken to identify major domains (components) within the CORE-OM and to measure the correlations of each item with underlying components. First, the Kaiser-Meyer-Olkin measure of sampling adequacy was used to test whether correlations between CORE-OM items can be explained by other underlying variables, and therefore to indicate whether PCA was appropriate for the analysis of CORE-OM data. A value of the measure closer to 1.0 indicates that PCA is an appropriate method of analysis (Cerny & Kaiser, 1977). In addition, Bartlett's test of sphericity was used to test the null hypothesis that the CORE-OM items are uncorrelated. A significant test confirms the appropriateness of PCA (Bartlett, 1950).

Significant components, i.e. components that mostly contribute to the explanation of variance in the items, were identified using Horn's parallel analysis (Horn, 1965), as recommended in the literature (Zwick & Velicer, 1986), although eigenvalues with a value ≥ 1 and the screeplot of the analysis were also inspected. Subsequently, the rotated component matrices showing the sizes of the loadings of each item to each extracted component were examined to assess correlations of every item with each of the main components of the instrument. Due to the indicated correlations between the CORE-OM's conceptual domains, two types of rotation were used: orthogonal (Varimax method with Keiser normalisation), which assumes that all components are uncorrelated to each other, and oblique (Promax method with Keiser normalisation), which allows for correlations between components. Use of both types of rotation and comparison of the results has been recommended in the literature (Kieffer, 1998). It has been suggested that if the differences between the results of the two types of rotation are negligible, then interpretation of findings can be based on the orthogonal rotation. However, if there are significant discrepancies between the results, then interpretation of the oblique rotation should be preferred (Kieffer, 1998). In every rotated solution, the correlation matrices between items and components were inspected. For the oblique rotation, the pattern matrix of unique relationships between items and components (uncontaminated by the overlap among factors) as well as the component correlation matrix were assessed. In all

matrices, loadings with coefficients ≥ |0.40| were considered to reveal strong correlations between an item and a component, and between components regarding the oblique rotation. Loading of items on the same component was considered as a strong indication that these belonged to the same underlying domain captured by the CORE-OM.

4.2.2 Step 2. Reduction of items and response levels from the CORE-OM

The suitability and performance of the CORE-OM items were mainly assessed by conducting Rasch analysis, supported by the results of classical psychometric tests. These techniques examined the psychometric properties of CORE-OM items and indicated which items should be considered for exclusion. The decision for omission of items from the new measure was determined by the interpretation of the results of the analyses conducted at this stage, and further judgments on the content of the items in the context of the eventual derivation of a PBM for common mental health problems. The final decision for exclusion of items from the health state classification was agreed with the thesis advisory group, which provided expert opinion. Further to item reduction, Rasch analysis provided a guide for the reduction in the response levels of the items included in the final measure.

Rasch analysis

Rasch analysis was used to assess the psychometric properties of the CORE-OM items and the optimal levels of response, and ultimately assist in the selection of best performing items for inclusion in the new PBM, in accordance with the methodology employed in published literature (Young et al., 2009 & 2011). From early stages of data analysis it was decided that if initial analyses of Step 1 provided further indications of high correlations between the CORE-OM domains and items, then Rasch analysis would not be conducted separately on each domain of CORE-OM, as described in relevant literature (Young et al., 2009 & 2011), because this methodology entailed the danger of generation of implausible health states, as already discussed. Instead, Rasch analysis would be undertaken on the whole CORE-OM instrument, in order to construct a unidimensional measure that fulfils the criteria of the Rasch model. The justification for this type of analysis lies in the properties of the Rasch

model, which allow conversion of respondents' ordinal scores into a continuous scale, allocation of respondents along this scale based on their symptom severity, and, consequently, identification of distinct groups of respondents with various levels of symptom severity, which translate into plausible health states, as discussed later in this chapter and reiterated in Chapter 6.

Results of statistical tests employed at the first step of analysis revealed that, indeed, CORE-OM does not have a clear multidimensional structure; the major domains of the instrument are highly correlated. The findings of these analyses are provided in detail in Chapter 5. These findings confirmed that the standard approaches for developing a health state classification and for generating health states for use in a valuation survey would not be appropriate in the case of CORE-OM, since they were likely to eventually lead to the generation of implausible health states. Based on these findings, it was decided to undertake Rasch analysis on the whole CORE-OM instrument, aiming at discarding items not fitting to the Rasch model and ultimately producing a unidimensional scale. As it is discussed in Chapter 5, re-scoring of items was necessary due to several items having disordered thresholds. Since no common re-scoring of all 34 items was possible to achieve, the partial credit Rasch model was used for the analysis.

Items not fitting in the Rasch model were excluded one at a time, followed by Rasch analysis on the remaining items and subsequent testing of fit statistics. The order of exclusion of items was based on expert opinion after considering the results of Rasch analyses and classical psychometric tests. This process was repeated until all remaining items fit in the Rasch model. The PSI was checked in each consecutive Rasch analysis to ensure that the model retained good ability to discriminate amongst different respondent groups. In addition, the class interval structure was inspected to confirm a homogeneous allocation of respondents across class intervals, which would be an indication that estimates of chi square statistics were reliable.

The following Rasch analysis criteria were considered in Step 2, in order to exclude non-fitting items and reduce response levels:

Item level (threshold) ordering and response level rescoring

Rasch item threshold maps were inspected to investigate whether the scoring categories of each item progressed in a logical order, that is, whether respondents were able to distinguish between adjacent response levels. Normally, persons with higher symptom severity are expected to obtain higher scores in each item, and this increase in scores alongside increases in symptom severity should happen systematically in a logical progression. When the probability of selecting higher (more difficult) response levels decreases with respondents' increased symptom severity, it is an indication that respondents have difficulty in understanding the differences between item response statements and are not able to discriminate between adjacent levels of response.

When an item had disordered thresholds (i.e. when an item score was likely to decrease as respondent's severity increased), then item rescoring was attempted, that is, adjacent response levels of this item were combined (merged), in order to achieve ordered thresholds. Several ways of rescoring were attempted. The following criteria were used in order to reduce response levels of items with disordered thresholds and achieve threshold ordering:

- Visual inspection of category probability curves for each item: these curves show the probability of a person selecting each response level of an item depending on their ability across the Rasch model logit scale. Curves that appeared to have considerable overlapping in the graph (in terms of the area under the curve) indicated that respondents had difficulty in distinguishing between respective (adjacent) response levels, and therefore such response levels were candidates for merging.

- Examination of category response proportions: response levels with very low proportion of responses were candidates for merging with an adjacent response level, as they did not add much information about the respondents' severity of symptoms. Nevertheless, this criterion was not

applied on the highest response levels of ‘difficult’ items (as indicated by their location), as even low proportions of responses in these cases indicated respondents with potentially very severe symptoms whom the new health state classification should be able to identify, since the aim was the development of a measure capturing the full range of symptom severity of common mental health problems.

- Reduction in the number of scoring categories should be balanced between developing a concise number of response levels and minimum loss of information

- The new (merged) response categories should be clinically meaningful without reducing significantly the information on the respondents’ symptom severity. Clinically meaningful combinations of response levels were considered to be:
    - ‘never’ – ‘only occasionally’
    - ‘only occasionally’ – ‘sometimes’
    - ‘often’ – ‘most or all the time’

  If the only way to order an item’s thresholds was by merging adjacent responses that were judged to be not clinically meaningful (such as ‘sometimes’ and ‘often’), then this item was a candidate for exclusion from the final measure. On the other hand, with regard to difficult items such as the risk items, it was decided that merging the adjacent response levels ‘never’ and ‘only occasionally’ was not appropriate, as in this case these indicated significantly different levels of symptom severity.

Following item rescoring attempts, subsequent Rasch analyses were conducted to confirm that all items had ordered thresholds.

Goodness of fit with the Rasch model after threshold ordering

After item rescoring and threshold ordering, overall fit statistics (item-person and item-trait interaction statistics) as well as individual item fit statistics were measured to assess to what extent the measure fit in the Rasch model. The fit of data into the model is indicated by a fit residual, which expresses the

difference between the persons' observed responses and those expected according to the Rasch model, and approximates a Z-standardised normal distribution. The item-trait interaction, which tests the overall fit between items and persons across the scale, was determined by a chi-squared probability. Regarding item-person interaction statistics, a fit residual of the mean of the distributions around zero and a fit residual of the standard deviation close to one indicated a good model fit. Regarding item fit residuals, those beyond ± 2.5 were considered to indicate a source of misfit in the model. Item fit residuals with a value below -2.5 indicated that items over-discriminated among respondents; such items were likely to summarise the rest of the items and therefore were redundant (i.e. they did not provide any extra information relative to the rest of the instrument); item fit residuals with values above +2.5 were signs of under-discrimination, and subsequently such items were also candidates for exclusion. Misfit of an item could indicate that the item is badly conceptualised, that it belongs to another domain and not to the underlying unidimensional scale, or that it cannot target well the study population (for example it is extremely easy or difficult). In addition to the fit residual, item fit was indicated by a chi-squared probability. Person fit residuals were checked in consecutive analyses to assess the level of outliers in the model. However, persons with 'extreme' fit residuals were not excluded from analysis, with the rationale that such persons are part of the study population, a realistic picture of which should be reflected in the results of Rasch analysis. Significance levels (probabilities) for the chi-square tests were calculated using Bonferroni adjustments, based on the number of CORE-OM items, to account for multiple testing (Bland & Altman, 1995).

Differential Item Functioning

All CORE-OM items were assessed for DIF. Items demonstrating significant and persistent DIF in consecutive analyses (that is, item responses depended on patients' demographic characteristics) were strong candidates for exclusion from the final PBM. This was decided for two reasons: first because DIF is a source of misfit in the Rasch model; and second, because items included in a PBM need ideally to constitute a universal measure, being perceived in a similar (and not systematically different) way by the whole patient population, as well as by the valuing population, regardless of their baseline

characteristics. For the same reason, although uniform DIF of an item can be dealt with by splitting the item for DIF and creating unique ‘sub-items’ corresponding to sub-groups differing in the baseline characteristic for which DIF was identified (Brodersen et al., 2007), this was not attempted in this analysis; therefore items demonstrating DIF were excluded from further consideration.

Based on the availability of baseline demographic data included in the CORE-OM dataset that was subject to Rasch analysis, items were examined for DIF on three demographic characteristics: gender, age and ethnicity. Gender was treated as a binary outcome. Regarding age, in order to create a categorical outcome, persons in the dataset were divided into the following age groups:

- Age ≤ 25 years
- Age between 26 and 40 years
- Age between 41 and 65 years
- Age > 65 years

Finally, in terms of their ethnicity, individuals in the dataset were categorised in the following sub-groups, after taking into account the percentage of people belonging to different ethnic groups that were included in the dataset (so that some ethnic groups with very low percentage of respondents were merged together):

- White
- Black
- Asian
- Other
- Mixed

Item location

Location of items was examined to assess their relative ‘difficulty’. In principle, the new health state descriptive system should include a range of items of varying difficulty, so that the respective PBM is able to target the study population well, capturing the whole range of symptom severity and distinguishing between different symptom severity levels.

Classical psychometric tests

Standard psychometric tests were performed as an extra tool in the assessment of psychometric properties of CORE-OM items. The results of these tests indicated less suitable items for inclusion in the final measure and were taken into account alongside the results of Rasch analysis when considering potential candidates for exclusion from Rasch analysis (and therefore from the final instrument). No strict thresholds were used to determine the performance of CORE-OM items in psychometric testing. The following psychometric criteria were considered:

Distribution of responses

Distribution of responses to each item was examined to explore the extent of floor or ceiling effects. Such effects would imply that an item does not efficiently target the severity of symptoms of the study population. In the context of Rasch analysis, ceiling effects (i.e. a large proportion of responses with a score 0 at baseline - for negatively worded outcomes) would indicate that an item is rather ‘easy’ and cannot capture the severity of symptoms met in the study population. Such an item is probably not appropriate for inclusion in a PBM (and indeed in any outcome measure). In contrast, items with floor effects (i.e. a large proportion of responses with a score 0 at baseline – for negatively worded items) are likely to comprise ‘difficult’ items that can identify more severe cases, the severity of which would not be possible to assess with an item of ‘average difficulty’. Therefore, items with floor effects are not considered to be unsuitable for inclusion in the final health state descriptive system, as long as they can identify a minimum number of more severe cases that is deemed to be significant.

Responsiveness

Responsiveness of each item was estimated by measuring the item change score between baseline and end-of-therapy. The measure of responsiveness was the SRM, defined as the change score of an item divided by the standard deviation of the change score. Although no cut-off points were used in order to judge the level of responsiveness, in general a SRM value of 0.2 to 0.3 was

deemed to indicate a small effect, a value around 0.5 a medium effect, and a value of 0.8 and above a large effect (Cohen, 1988).

In addition, because the responsiveness of an item can be affected by floor or ceiling effects, it was decided to conduct a sub-analysis and measure responsiveness for each item only for persons who had provided responses to this particular item with scores ranging from 1 to 4 (“only occasionally” to “most or all the time” for a negatively worded item) at baseline, thus excluding respondents with a baseline score of 0 to this item. This was deemed useful because some of the more difficult CORE-OM items (for example item 16, I made plans to end my life) express more severe symptoms and are likely to demonstrate floor effects, with the majority of the study population responding “not at all” at baseline. Therefore, the average responsiveness for these items is expected to be low, since in the majority of the study population there is no scope for improvement. However, such items may be very useful in identifying people with severe symptomatology and assessing their responsiveness to treatment, and therefore should be still considered for inclusion in the final PBM, despite of demonstrating low overall responsiveness due to floor effects at baseline.

Correlation of each item score with the total CORE-OM score

This test was undertaken to explore the degree to which each item measures the same attribute with that measured by the whole questionnaire. Items with high correlation with the CORE-OM are judged to be good representatives of the whole instrument and good candidates for a concise PBM that aims to summarise information from a larger instrument with minimum information loss. Correlation was expressed using Spearman’s non-parametric p values.

Percentage of missing data

The percentage of missing data for each item was measured as an estimate of the item’s acceptability to the study population. High non-response rates imply also a difficulty in understanding the item which reduces its usefulness as part of a health state descriptive system.

4.2.3 Step 3. Selection of CORE-OM items for inclusion in the health state classification

After items that were judged to be inappropriate or less suitable for inclusion in the final measure were excluded from further analysis and a unidimensional scale fitting the Rasch model was constructed, further reduction of items was attempted. This was essential as the purpose of this process was to develop a concise PBM that is manageable in a valuation survey with, nonetheless, minimum loss of information relative to the original measure; evidence has shown that respondents can receive, process, and remember about 7 ± 2 pieces of information, depending on the complexity of the statements (Miller, 1956). In terms of statistical methods used, the final selection of items for inclusion in the new PBM was based exclusively on Rasch analysis. In order to make the final item selection, different combinations of the remaining (fitting) items were tested against the following criteria:

- **Wide coverage of CORE-OM domains**
  The final instrument should consist of items representing the various domains of the CORE-OM, either expressed by the conceptual domains of the CORE-OM or as indicated by PCA (if different). Items in the final health state classification should express the maximum possible number of these domains, so that the new PBM is able to tap a range of different aspects of HRQoL that are relevant to people with common mental health problems, as captured by the CORE-OM, with minimum loss of information.

- **Best model and individual item fit**
  Overall model and individual item fit statistics should demonstrate best possible fit of the measure into the Rasch model.

- **Consistency in response levels across items**
  Response levels should ideally be the same for all items and reflect clinically meaningful situations; consistency of scoring categories across items included in the final instrument was attempted for practical purposes, as it was considered that participants in a valuation survey were likely to better understand and value items with the same response levels. This

criterion meant that some items could potentially be rescored, despite already having ordered thresholds, in order to achieve consistency regarding response statements across all items in the final measure.

- **Coverage of the full range of symptom severity**
  The final instrument should be well targeted and able to capture the whole range of symptom severity observed in the study population; in order to achieve this, items should cover different locations across the latent variable. In a well-targeted instrument, the average location of the study population is expected to coincide with the average location of the items. Moreover, the extent of targeting can be assessed by inspection of the item map, which displays the person-item targeting distributions.

- **Reliability**
  The final measure should have acceptable reliability, expressed by the PSI, having in mind that the ability of the measure to discriminate amongst different respondent groups would likely need to be traded off with its conciseness and convenience in using as a PBM. PSI is expected to be greatly reduced with significant reduction in the number of items. Generally, a PSI of 0.7 is regarded as the lowest acceptable level of reliability (Fisher, 1992).

- **Unidimensionality**
  This property was tested using an extra post-hoc test (Smith, 2002) as recommended in the literature (Tennant & Conaghan, 2007; Tennant & Pallant, 2006). The first stage of this test was to undertake PCA on the item fit residuals in order to identify the first residual factor that primarily contributes to the variance of data after the 'Rasch factor' has been accounted for. Subsequently, the nature of the correlation between the items and the first residual factor was examined, in order to define two subsets of items, those positively and those negatively correlated with the first residual factor. These two 'divergent' sets of items, which were most likely to breach the assumption of unidimensionality, were used to estimate two separate scores for each respondent, respectively. If the content of the

whole scale was unidimensional, then each respondent should produce similar scores in the two subsets. Thus, independent t-tests were 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. According to the post-hoc test, if the proportion of independent *t*-tests fell outside the boundaries of acceptable significance, this would be an indication that there might still be some degree of multidimensionality within the whole construct, as respondents would be shown to behave differently in each of the 2 (divergent) subsets of items. If the proportion of significant independent *t*-tests was lower than 5%, this would confirm the unidimensionality of the scale; if the proportion of significant independent tests exceeded 5%, then a binomial confidence interval for proportions would need to be estimated: a lower 95% confidence interval below 5% would be an indication of unidimensionality.

- **Local item independence**
  This was confirmed by checking the Varimax Rotation loadings produced by the PCA on the item fit residuals: if local independence held, each item of the final scale should load highly on separate residual components each, indicating that no item is highly correlated with the others. Moreover, the correlation of residuals between pairs of items was examined in the residual correlation matrix: correlations between residuals within ± 0.40 were an indication of local item independence.

In addition to the above criteria, an extra criterion was set for the construction of the final measure. This was directly related to the ability of the Rasch model to assign persons on the Rasch model logit scale based on their responses, thus generating groups of respondents of different symptom severity (Bond & Fox, 2007) corresponding to plausible health states:

- **Increase in item threshold locations with increasing difficulty of the item**

  Thresholds are the points on the logit scale where the probability of a response to two adjacent response levels, like 0 and 1, or 1 and 2, etc. is equally likely (50%). The difficulty of an item is expressed by its average location on the scale. According to this criterion, for every item, the threshold location of any two adjacent response levels (for example of levels 0 and 1) should increase as the difficulty of the item increases. So the threshold location of response levels 0 and 1 should be lower for an easier item compared with a more difficult one; similarly, the threshold location of response levels 1 and 2 should also be lower for the easier item. This condition should apply across all thresholds and all items on the final scale. Threshold locations for all adjacent response levels of all items were compared by visual inspection of the item threshold map; this is an output of Rasch analysis that depicts the most likely item response combinations expected for each location across the Rasch model logit scale. Fulfilment of this criterion ensured a ‘smooth’ transition of responses from milder to more severe health states, allowing clear depicting of plausible health states on the Rasch item threshold map. Identification and selection of plausible health states was crucial for the valuation of the new instrument, as illustrated in Chapter 6 (section 6.2.1).

The combination of items that met as many of the above set criteria as possible formed the final health state classification that led, following valuation, to the development of a decision-specific PBM for people with common mental health problems.

4.2.4 Step 4. Validation of the new health state classification

The new health state classification was validated by repeating the above process on three different samples of respondents. In both of those samples, the final measure was tested for overall and item fit statistics, DIF, reliability, targeting of study population and local independence of items. The post hoc unidimensionality test was repeated and the Rasch item threshold map was inspected to confirm the smooth transition of responses from milder to more

severe health states and that the same plausible health states were identified in all samples.

PCA and classical psychometric tests were conducted in SPSS 19 (IBM Corp., 2010). Monte Carlo PCA for Parallel Analysis software was used to identify significant eigenvalues for PCA according to Horn's parallel analysis (Watkins, 2008). Rasch analysis was undertaken in RUMM2020 (Andrich et al., 2003).

4.2.5 Expert opinion – other considerations

A thesis advisory group was set up at the start of the development of the new PBM in order to advise on the appropriateness and suitability of each of the CORE-OM items for inclusion in a PBM for common mental health problems, following interpretation of the results of the statistical analyses and further considerations regarding the relevance of some of the items. Interpretation of the findings relied on many occasions on the group's judgment rather than pre-determined psychometric cut-off points, as these might have statistical but not necessarily clinical relevance.

The group met at regular intervals to review the results of the analyses undertaken up to that point and subsequently advise on further steps. Alongside the results of the statistical analyses, the group's judgments contributed significantly to the final decisions regarding the inclusion or exclusion of certain CORE-OM items and the construction of the final PBM.

The thesis advisory group consisted of the following members:

- Professor John Brazier, who provided expert opinion on the overall methodology to be used for the derivation of a health state classification from the CORE-OM, the interpretation of the results of statistical analyses and the suitability of items for inclusion in the new PBM, given his involvement in the derivation of numerous PBMs from existing generic and condition-specific non-PBMs, including the derivation of the SF-6D from the SF-36 (Brazier et al., 2002).

- Professor Michael Barkham, one of the developers of the CORE-OM and a CORE System Trustee, who provided background information on the

original instrument and advised at various stages of the development of the new measure regarding the structure and the conceptual domains of CORE-OM, the appropriateness and relevance of each of the CORE-OM items for inclusion in a PBM, as well as the interpretation of the findings of the statistical analyses, given his involvement in both the development and the application of the CORE-OM in clinical practice and research (Barkham et al., 2001; Evans et al., 2000; Stiles et al., 2008).

- Dr Tracey Young, who advised on the methods to be adopted for the development of the new measure, the interpretation of the findings of the statistical analyses and the suitability of items for inclusion in the new PBM, as she has been involved in the derivation of several PBMs from existing CSMs, using mainly Rasch analysis and classical psychometric tests (Young et al., 2009 & 2011).

4.3 CORE-OM datasets used in the analyses

Data analysed in order to construct a health state classification from the CORE-OM were derived from a database service containing information on 6,610 clients from 33 NHS primary care counselling services. The database was created through the accumulation of data from a data mounting, analysis and reporting service based at the Psychological Therapies Research Centre (PTRC), University of Leeds. Counselling services sent completed batches of CORE system forms to PTRC for analysis and reporting. Data mounting was automated by the FormicTM system which exports the data in SPSS data files which are then checked thoroughly for scanning and data entry errors. Services agreed to the accumulation of anonymous data into a cumulative database. Each service was given a 3-hour training session which included an introduction to the system, its rationale and advice on completion of the forms. In addition, each participating practitioner was provided with a comprehensive user manual that contained scoring information and guidelines for completion of the CORE system measures (CORE System Group, 1999). Services also had telephone support from the CORE team to deal with specific queries if required. Details on the full dataset and data collection procedures are available in Evans and colleagues (2003). A random sample of 1,500 primary

care clients from this database formed the dataset analysed for the purposes of this thesis [N1500].

The dataset included data on clients' demographic parameters, information on client history and assessment outcome, service parameters relating to therapy, as well as clients' scores on each of the 34 CORE-OM items at baseline and end of therapy. All variables contained in the dataset are provided in Table 9.

PCA was undertaken on the baseline data of the whole CORE-OM dataset [N1500] and repeated on each of two random sub-sets [N750a] and [N750b] which the whole dataset [N1500] was split into, in order to test the reproducibility of the PCA findings on [N1500].

Regarding Rasch analysis, it has been shown that a number of statistics for polytomous scales (such as chi square statistics) are highly dependent on the sample size used (Smith et al., 2008). It has been argued that analyses of large sample sizes can demonstrate misfit even if data actually fit the model; this occurs because fit statistics become more powerful as the sample size increases, and with large sample sizes, even the slightest misfit will be exposed, translating into a higher probability for type I errors with increased sample size (Linacre, 2003). Rasch analysis was thus performed on a sub-sample of 400 randomly selected respondents [N400a] out of the [N1500] dataset.

Standard psychometric tests contributing to the assessment of psychometric properties of CORE-OM items were also performed on the sample [N400a], which was used in Rasch analysis. This was decided because results of psychometric tests supplemented the results of Rasch analysis and it was deemed more appropriate for results of both types of analysis to refer to the same study sample.

Table 9. Demographic, history, assessment and other types of data contained in the dataset [N1500] analysed in this thesis in order to derive a health state classification from the CORE-OM

- Demographic parameters
  - Gender
  - Ethnicity
  - Age
  - Employment status
  - Place of residence, relationships and support
- History
  - Previously seen for therapy
  - First or follow-up assessment
  - Number of previous episodes
  - Months since last episode
  - Concurrent or previous primary, secondary and specialist care
  - Prescribed medication
  - Problem mix (severity and duration) assessed by the practitioner
    - ✓ Depression
    - ✓ Anxiety/stress
    - ✓ Trauma/abuse
    - ✓ Bereavement/loss
    - ✓ Psychosis
    - ✓ Self esteem
    - ✓ Personality problems
    - ✓ Interpersonal/relationship problems
    - ✓ Cognitive/learning problems
    - ✓ Living/welfare problems
    - ✓ Eating disorder
    - ✓ Work/academic problems
    - ✓ Physical problems
    - ✓ Addictions
    - ✓ Suicide risk
    - ✓ Self harm risk
    - ✓ Harm to others risk
    - ✓ Legal/forensic problems
  - ICD-10 diagnosis
  - Negative and positive actions to cope with problems
- Baseline scores on 34 CORE-OM items
- Assessment outcome
  - Problem resolved
  - Accepted for further sessions
  - Referred back to referrer or to other service
  - Therapy declined by the client
- Service parameters
  - Mean waiting time to first appointment (days)
  - Mean number of sessions offered / attended
  - Type of therapy
  - Frequency of therapy
  - Discontinuation and reasons
- Therapy outcome
  - Problem mix (severity and duration) assessed by the practitioner
  - End-of-therapy scores on 34 CORE-OM items
  - Other benefits
    - ✓ Coping
    - ✓ Subjective well-being
    - ✓ Day-to-day functioning
    - ✓ Personal relationships

The final measure was validated on three separate study samples:

a. another random sub-sample of 400 respondents [N400b] out of the [N1500]
b. the whole initial sample [N1500] after adjusting the sample size for use in the test-of-fit statistics
c. because the [N1500] dataset consisted of patients presenting to primary care services and there were concerns that the newly developed instrument might not be representative of the intended study population (that is, the whole population of people with common mental health problems irrespective of their level of severity or their site of access), the results of the final solution of Rasch analysis were also validated on a separate 'mixed' sample of 1,500 patients attending either primary or secondary care [N1500v]. This sample was randomly selected from a dataset of 7,651 people with common mental health problems; data were collected from 49 NHS sites routinely using the CORE-OM to monitor patients at intake. These sites comprised counselling or psychology services within primary care groups or Trusts, or secondary care settings providing clinical psychology and psychotherapy services. The dataset, which is described by Barkham and colleagues (2005b), was available from the same data mounting, analysis and reporting service based at the PTRC, University of Leeds, that provided the [N1500] dataset. Baseline demographic, history and CORE-OM data were available for 1,390 persons in the [N1500v] dataset. Rasch analysis on the [N1500v] was also adjusted for sample size for use in the test-of-fit statistics, to avoid the risk for type I errors due to large sample size. Further to the validation of the final solution of Rasch analysis, [N1500v] was used to explore potential presence of DIF in the final solution regarding the site of patient access.

4.4 Summary

This chapter presented the methods that were employed in order to derive a health state classification from the CORE-OM. The methodology adopted included techniques such as PCA, Rasch analysis and classical psychometric tests, which have been widely used for the derivation of multidimensional health state classifications from existing measures in the literature. The methodology that was followed in the case of CORE-OM was different from that reported in the relevant literature, dictated by the measure's lack of clear multidimensionality and the danger of generating implausible health states if the 'standard' health state classification approach and statistical techniques for generation of health states were to be followed. Chapter 5 presents the results of the analyses of all methodological steps proposed in this chapter that led to the development of a new health state classification for people with common mental health problems.

Chapter 5. Results on the derivation of a health state classification from the CORE-OM

5.1 Introduction

This chapter presents the results of all the analyses undertaken on the CORE-OM including PCA, Rasch analysis and classical psychometric tests, together with the considerations and expert advice of the thesis advisory group, that led to the derivation of a new health state classification system for people with common mental health problems that is amenable to valuation. It follows from the 4-step proposed methodology outlined in the previous chapter. The results presented here have also been reported in a paper publication (Mavranezouli et al., 2011).

5.2 Characteristics of the study sample

The [N1500] CORE-OM dataset that was used for PCA included 1500 people with common mental health problems presenting to NHS primary care counselling services; baseline demographic, history and CORE-OM data were available for 1320 persons. A random sub-sample of 400 cases out of the [N1500] was used in Rasch analysis and classical psychometric testing [N400a]. A summary of the baseline demographic and history characteristics of the study sample [N1500] as well as of the sub-sample [N400a] are shown in Table 10.

Table 10. Demographic and history characteristics of the study sample [N1500] and the random sub-sample [N400a] analysed in this thesis in order to derive a new health state classification from the CORE-OM

Parameter	[N1500]	[N400a]
Mean age (standard deviation)	38.4 (13.3)	38.2 (13.4)
Age distribution		
≤ 25	16.4%	16.4%
26 - 40	46.3%	43.6%
41 - 65	32.9%	36.2%
> 65	4.4%	3.8%
Gender - female	72.0%	71.5%
Ethnicity		
Asian	4.2%	3.2%
Black	1.9%	1.8%
White	92.2%	92.6%
Other - Mixed	1.7%	2.4%
Employment		
Full-time paid employment	39.8%	40.1%
Part-time paid employment	14.4%	12.5%
Receiving sickness benefit	7.7%	7.1%
Unemployed	13.1%	10.9%
Full-time student	2.5%	2.7%
Part-time student	1.8%	1.8%
Retired	5.8%	6.1%
House person	13.5%	17.3%
Other	1.4%	1.5%
Place of residence/relationships/support		
Living alone (not including dependents)	29.8%	30.2%
Living with partner	50.7%	51.1%
Living with parents/guardian	9.4%	9.6%
Living with other relatives/friends	6.7%	5.6%
Living in shared accommodation	2.5%	2.7%
Living in temporary accommodation	0.8%	0.7%
Living in institution/hospital	0.1%	0.0%
Caring for a child	37.8%	39.0%
Full time carer	1.0%	1.4%
Other service use for psychological problems		
Primary care		
Concurrent - Previous	41.7% - 15.1%	40.3% - 14.5%
Secondary care (any setting)		
Concurrent - Previous	2.7% - 7.9%	3.0% - 5.6%
Specialist care (any setting)		
Concurrent – Previous	0.9% - 5.7%	0.8% - 4.3%
Mean CORE-OM score at intake (standard deviation)	18.17 (6.75)	18.11 (6.50)

The [N1500] consisted mainly of women (72.0%) and white population (92.2%). The mean age of the sample was 38.4 years (range 12 to 88 years, standard deviation 13.3 years). The majority was in full-time (39.8%) or part-time (14.4%) paid employment, while 13.1% were unemployed and 13.5% considered themselves as house persons. A small proportion (5.8%) of the sample was retired, and another 4.3% was full- or part-time students. Just over 50% of the sample lived with a partner, 29.8% lived alone, and 37.8% looked after at least one child. A large proportion of the sample stated they received other services for a psychological problem at the time of presentation: 41.7% in a primary care setting, 2.7% in secondary care and 0.9% in specialist care. Moreover, a proportion of the sample stated they had received services for a psychological problem in the past (15.1%, 7.9% and 5.7% in primary, secondary and specialist care, respectively). The average CORE-OM score was 18.17 (standard deviation 6.75) on a scale 0-40. The demographic and history characteristics were very similar for the sub-sample [N400a].

5.3 Results of Step 1: exploration of the dimensionality of the CORE-OM

5.3.1 Principal Component Analysis

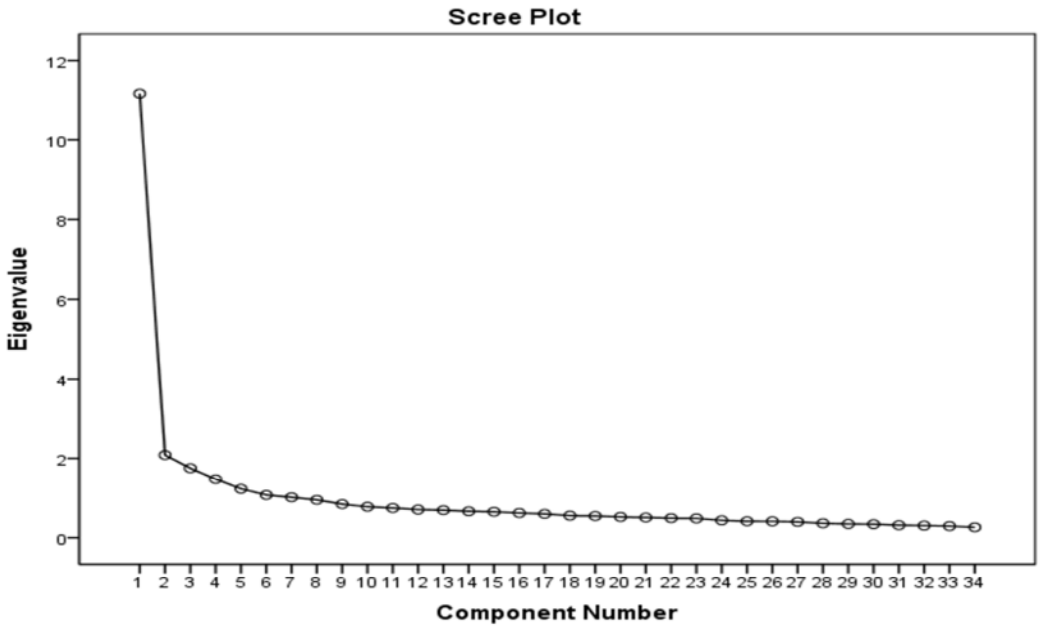
PCA was undertaken on [N1500], as well as on two random subsets [N750a] and [N750b] the study sample was split into. Analysis of the CORE-OM data in [N1500] showed that the Kaiser-Meyer-Olkin measure of sampling adequacy reached 0.95, meaning that factoring of data was appropriate and meaningful. Bartlett's test of sphericity demonstrated the significance of the findings ($p<0.001$). Although the analysis identified 7 components with eigenvalues above 1, Horn's parallel analysis indicated 5 significant components (Table 11). The latter finding is in agreement with the screeplot of the analysis, which is provided in Figure 7: the slope of the line appears to change after the 5th component, suggesting that only the first 5 components are significant.

Table 11. Significant components of CORE-OM identified by Principal Components Analysis in [N1500] - Comparison of components with eigenvalues >1 with significant components identified by Horn's parallel analysis

Component	PCA: Initial Eigenvalues			Horn's parallel analysis: Significant mean eigenvalues (SD)	
	Total	% of Variance	Cumulative %		
1	**11.15**	32.8	32.8	**1.29**	(0.02)
2	**2.08**	6.1	38.9	**1.26**	(0.02)
3	**1.75**	5.2	44.1	**1.23**	(0.01)
4	**1.48**	4.3	48.4	**1.21**	(0.01)
5	**1.23**	3.6	52.0	**1.19**	(0.01)
6	**1.07**	3.2	55.2	1.17	(0.01)
7	**1.02**	3.0	58.2	1.15	(0.01)

Significant eigenvalue levels identified using each approach are provided in bold; SD = standard deviation

Figure 7. Screeplot of Principal Component Analysis in [N1500]



Loadings of the CORE-OM items on the identified significant components in the whole dataset [N1500] are presented in Table 12, which depicts the rotated component matrix resulting from orthogonal rotation, and in Table 13, which provides the pattern matrix of oblique rotation, that is, the matrix that allows identification of the unique relationships between items and components, uncontaminated by the overlap across components. Strong loadings were considered those with a correlation coefficient $\geq |0.40|$. Results were very similar between the two methods of rotation. The majority of CORE-OM items loaded on the same component(s) regardless of the method of rotation; the exception to this pattern were 5 items (items 2, 15, 17, 23, 27) that were shown to load on component 1 and/or component 2 in orthogonal rotation but not in oblique, as well as item 31 that was shown to load on components 2 and 5 in orthogonal rotation but only on component 1 in oblique.

Table 12. Findings of Principal Components Analysis on CORE-OM data in [N1500]. Orthogonal rotation – rotated component matrix

CORE-OM items	Components						
	1	2	3	4	5	6	7
1. [terribly alone and isolated]	**.50**	.39	.28	.18	.20	-.07	.11
2. [tense, anxious or nervous]	**.42**	**.46**	.22	-.00	-.12	.02	.38
3. [somebody to turn to for support]	.16	.09	.22	.04	**.71**	.04	.07
4. [felt ok about myself]	.32	**.56**	.17	.14	.30	-.01	-.02
5. [totally lacking in energy and enthusiasm]	.40	**.46**	.06	.06	.06	.05	.29
6. [physically violent to others]	.03	.12	.02	.18	-.03	**.80**	.02
7. [able to cope when things go wrong]	.29	**.65**	.06	.15	.08	.10	-.08
8. [aches, pains, physical problems]	.13	-.00	-.01	.05	.11	-.01	**.81**
9. [thought of hurting myself]	.18	.17	.11	**.83**	.06	.08	.02
10. [talking to people has felt too much]	.30	.27	.22	.07	.27	.07	.33
11. [tension/anxiety prevented doing things]	.22	**.60**	.23	.11	-.10	.10	.41
12. [happy with the things I've done]	.22	**.65**	.18	.07	.29	.08	-.09
13. [disturbed by unwanted thoughts/feelings]	**.61**	.09	.16	.22	-.01	-.03	.22
14. [felt like crying]	**.70**	.27	.13	.12	.03	.05	-.06
15. [felt panic or terror]	.37	**.40**	.16	.21	-.22	-.03	.36
16. [made plans to end my life]	.11	.13	.10	**.82**	.01	.02	.06
17. [overwhelmed by my problems]	**.58**	**.48**	.23	.12	-.01	.06	.10
18. [difficulty of getting to sleep/staying asleep]	**.60**	.13	-.08	.07	.20	.10	.21
19. [felt warmth or affection for someone]	-.05	.24	.02	.12	**.64**	.05	.02
20. [problems impossible to put to one side]	**.61**	.37	.20	.03	-.01	.09	.06
21. [able to do most things I needed to]	.08	**.72**	.10	.11	.07	.14	.12
22. [threatened or intimidated another person]	.07	.04	.24	.07	.05	**.75**	-.01
23. [felt despairing or hopeless]	**.48**	**.48**	.34	.21	.09	.04	.12
24. [thought it would be better if I were dead]	.28	.22	.29	**.69**	.10	.04	.01
25. [felt criticised by other people]	.19	.16	**.76**	.06	.08	.17	.06
26. [thought I have no friends]	.20	.15	**.61**	.17	.31	.07	.08
27. [felt unhappy]	**.64**	**.42**	.22	.10	.18	.02	-.03
28. [unwanted images/memories distressing]	**.68**	.02	.20	.22	.02	.05	.07
29. [irritable when with other people]	**.46**	.14	.30	.00	.17	.39	.04
30. [I am to blame for problems & difficulties]	.28	.24	**.47**	.18	.14	-.00	-.13
31. [felt optimistic about my future]	.15	**.50**	.01	.05	**.41**	-.12	.05
32. [achieved the things I wanted to]	.14	**.61**	.17	.13	.31	.05	.01
33. [felt humiliated or shamed by other people]	.17	.16	**.74**	.17	-.03	.15	.09
34. [hurt myself physically/risks with health]	.10	.02	.04	**.59**	.15	.33	.06

Rotation method Varimax with Kaiser normalisation. Loadings ≥ │0.40│ are shown in bold; loadings on the 2 components that were found to be non-significant by Horn's analysis have been shaded in grey.

Table 13. Findings of Principal Components Analysis on CORE-OM data in [N1500]. Oblique rotation - pattern matrix

CORE-OM items	Components						
	1	2	3	4	5	6	7
1. [terribly alone and isolated]	**.42**	.22	.15	.06	.12	-.14	.03
2. [tense, anxious or nervous]	.26	.37	.11	-.13	-.17	-.02	.29
3. [somebody to turn to for support]	.09	-.05	.18	-.04	**.72**	-.01	.15
4. [felt ok about myself]	.16	**.57**	.01	.03	.21	-.06	-.09
5. [totally lacking in energy and enthusiasm]	.30	**.42**	-.13	-.05	.01	.02	.22
6. [physically violent to others]	-.02	.18	-.10	.12	-.05	**.81**	-.01
7. [able to cope when things go wrong]	.13	**.75**	-.14	.05	-.03	.07	-.20
8. [aches, pains, physical problems]	.01	-.16	-.05	.02	.21	-.03	**.90**
9. [thought of hurting myself]	.04	.06	-.04	**.85**	-.00	.02	-.02
10. [talking to people has felt too much]	.17	.14	.13	-.04	.25	.02	.33
11. [tension/anxiety prevented doing things]	-.12	**.63**	.14	.00	-.15	.06	.31
12. [happy with the things I've done]	.01	**.75**	.03	-.05	.19	.04	-.18
13. [disturbed by unwanted thoughts/feelings]	**.71**	-.20	.01	.13	-.06	-.08	.15
14. [felt like crying]	**.85**	.06	-.09	-.01	-.08	.01	-.18
15. [felt panic or terror]	.21	.32	.04	.13	-.28	-.07	.25
16. [made plans to end my life]	-.05	.02	-.00	**.86**	-.04	-.04	.03
17. [overwhelmed by my problems]	**.52**	.35	.05	-.02	-.12	.00	-.04
18. [difficulty of getting to sleep/staying asleep]	**.79**	-.08	-.32	-.02	.17	.08	.19
19. [felt warmth or affection for someone]	-.19	.30	-.08	.08	**.64**	.03	.10
20. [problems impossible to put to one side]	**.64**	.22	.02	-.11	-.11	.05	-.06
21. [able to do most things I needed to]	-.25	**.91**	-.05	.02	-.01	.12	.02
22. [threatened or intimidated another person]	.02	-.00	.21	-.02	.02	**.73**	-.04
23. [felt despairing or hopeless]	.32	.35	.21	.07	-.01	-.03	.01
24. [thought it would be better if I were dead]	.13	.04	.19	**.66**	.02	-.05	-.05
25. [felt criticised by other people]	-.08	-.05	**.91**	-.09	.01	.07	-.01
26. [thought I have no friends]	-.02	-.06	**.69**	.04	.26	-.03	.06
27. [felt unhappy]	**.66**	.26	.02	-.04	.06	-.04	-.14
28. [unwanted images/memories distressing]	**.86**	-.31	.04	.12	-.05	-.00	-.01
29. [irritable when with other people]	**.50**	-.06	.20	-.15	.11	.35	-.01
30. [I am to blame for problems & difficulties]	.14	.10	**.48**	.08	.04	-.08	-.21
31. [felt optimistic about my future]	-.01	**.59**	-.14	-.02	.37	-.15	.04
32. [achieved the things I wanted to]	-.13	**.71**	.04	.03	.23	.01	-.05
33. [felt humiliated or shamed by other people]	-.13	-.04	**.89**	.05	-.10	.05	.01
34. [hurt myself physically/risks with health]	.05	-.08	-.09	**.59**	.13	.30	.07

Rotation method Promax with Kaiser normalisation. Loadings ≥ │0.40│ are shown in bold; loadings on the 2 components that were found to be non-significant by Horn's analysis have been shaded in grey.

Table 14 provides a summary of the findings, showing the CORE-OM items that load on each of the underlying components (determined using a cut-off point of ≥ │0.40│for 'strong' loadings), and the conceptual domain of CORE-OM they belong to. Overall, the majority of the items loaded on the first two components (11 and 13 on each component, respectively, and 20 in total, using orthogonal rotation; 9 and 8 on each component, respectively, and 17 in total, using oblique rotation). Some items (2, 17, 23 and 27) loaded on both components under orthogonal rotation, but this finding was not supported in oblique rotation. It must be noted that items loading on the first 2 components cover 3 out of the 4 conceptual domains of CORE-OM, including symptoms/problems, functioning and subjective well-being. This implies that the conceptual domains of CORE-OM are not consistent with the components identified in this analysis. The results of PCA suggest that, in general, component 1 includes a number of items that belong to the conceptual domains of subjective well-being, symptoms and functioning, while component 2 covers positively worded items, together with some items representing subjective well-being, symptoms and functioning. Component 3 included the same 4 items in both rotations, mostly expressing functioning – close and social relationships, although one item (item 30) reflected symptoms - depression. Component 4, in both rotations, included the 4 items that comprise the risk/harm-to-self conceptual domain of CORE-OM. Component 5 included 3 items only, 2 items on functioning - close relationships (items 3 and 19) and item 31 on subjective well-being – the latter did not load on component 5 under oblique rotation. Component 6, which was non-significant according to Horn's analysis, included the 2 risk/harm-to-others items. Finally, the non-significant component 7 included item 8 (I have been troubled by aches, pains or other physical problems), which expresses physical health.

Table 14. Summary of findings of Principal Components Analysis in [N1500]. CORE-OM items with significant loadings (coefficients ≥ │0.40│) on underlying components – results of both orthogonal and oblique rotations

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

Items marked with asterisk (*) loaded strongly on the respective component only in orthogonal rotation; grey-shaded components are those found non-significant according to Horn's parallel analysis

The component correlation matrix in Table 15 shows the correlations between significant components, as indicated by oblique rotation. Loadings demonstrate a strong correlation between components 1, 2 and 3, and a moderate to strong correlation of component 4 with components 1 and 3.

Table 15. Findings of Principal Components Analysis of CORE-OM data in [N1500]. Oblique rotation – component correlation matrix

Component	1	2	3	4	5	6	7
1	**1.00**	**0.69**	**0.65**	**0.40**	0.15	0.10	0.36
2	**0.69**	**1.00**	**0.58**	0.36	0.17	0.06	0.35
3	**0.65**	**0.58**	**1.00**	0.**40**	0.18	0.20	0.25
4	**0.40**	0.36	**0.40**	**1.00**	0.12	0.16	0.15
5	0.15	0.17	0.18	0.12	**1.00**	0.06	-0.20
6	0.10	0.06	0.20	0.16	0.06	**1.00**	0.04
7	0.36	0.35	0.25	0.15	-0.20	0.04	**1.00**

Rotation method Promax with Kaiser normalisation. Correlation coefficients ≥ │0.40│ are shown in bold; correlations of the 2 components that were found to be non-significant by Horn's analysis have been shaded in grey.

Results of PCA on the two random sub-samples [N750a] and [N750b] are presented in Appendix 5. Findings were overall quite similar to those derived from analysis on the whole sample [N1500], in terms of the large number of items loading on the first two components and the correlations between domains identified by oblique rotation. More specifically, PCA on [N750a] identified 5 significant components according to Horn's parallel analysis (6 with eigenvalues above 1), which did not substantially differ between orthogonal and oblique rotation. The first two components included the majority of CORE-OM items (20 in total under orthogonal rotation and 19 under oblique rotation), with component 1 capturing mainly items relating to symptoms of emotional distress and component 2 containing the vast majority of the positively worded items. Component 3 appeared to cover items relating to functioning – close and social relationships, while components 4 and 5 covered risk/harm-to-self items and risk/harm-to-others items, respectively. Non-significant component 6 according to Horn's parallel analysis covered the physical item 8 (I have been troubled by aches, pains or other physical problems). The component correlation matrix produced by oblique rotation showed that component 1

correlated with all other significant components; however, no strong correlations between the other components were identified.

PCA on [N750b] revealed only 4 significant components based on Horn's parallel analysis (7 with eigenvalues above 1). There were more differences between the findings of orthogonal and oblique rotation than those observed between the two types of rotation in PCA in samples [N1500] and [N750a]. Under orthogonal rotation, 18 items loaded on the first two components, while under oblique rotation the respective number was 15 items. Component 1, as with other analyses, included items relating to symptoms of emotional distress, and component 2 included most of the positively worded items. However, some items relating to symptoms of emotional distress, which had been found to load on component 1 in previous analyses, were now found to load on component 3, together with item 8 (I have been troubled by aches, pains or other physical problems). Component 4 included all risk/harm-to-self items under both types of rotation. The last 3 components that were found to be non-significant according to Horn's parallel analysis included items on functioning – close and social relationships (component 5), items 3 and 19 on close relationships (component 6) and the two items expressing risk/harm-to-others (component 7). The component correlation matrix resulting from oblique rotation showed that the first 3 components were strongly correlated. In addition, non-significant component 5 was also strongly correlated with the first 3 components.

5.3.2 Summary and interpretation of findings – advisory group's views and decision on the approach to be adopted

The findings of PCA indicated that the CORE-OM consists of a large pool of items that belong to domains that are highly correlated; few items appear to belong to independent domains. The first 2 components identified in PCA contained the majority of CORE-OM items, with one component including mainly items expressing various symptoms of emotional distress and the other component made up mostly of positively worded items. A small number of items (items 25, 26, 29, 30 and 33) appeared to form a rather distinct group; with the exception of item 30, which conceptually belongs to the domain

symptoms – depression, the rest of these items conceptually express functioning – close or social relationships. Risk/harm-to-self and risk/harm-to-others items appeared to form two separate components, respectively. Items 3 and 19, which conceptually belong to functioning – close relationships, also formed a separate component. Finally, item 8 appeared to behave independently from the other CORE-OM items.

The loadings of the items on the identified components in the rotated matrices showed a similar pattern in both orthogonal and oblique rotations of PCA, indicating that orthogonal rotation might be an acceptable solution. On the other hand, oblique rotation revealed strong correlations between underlying domains identified by PCA. It is noticeable that the domains identified in PCA were overall not consistent with the conceptual structure of CORE-OM, as the first two components included items from all conceptual domains of CORE-OM, with the exception of risk items.

The findings of PCA performed in Step 1 suggest that the CORE-OM comprises a measure with no clear multidimensionality, since its domains (as identified by PCA) are highly correlated, perhaps with the exception of the risk items and items 8, 3 and 19, which overall appear to belong to separate, independent domains. It has to be noted, though, that literature suggests that it is possible that items with different levels of difficulty (e.g. items that capture different severity levels of mental symptoms) may form separate components in PCA, even though they may capture the same dimension; this may occur because ‘easy’ items (i.e. items capturing milder levels of disease) and ‘difficult’ items (i.e. items with an ability to identify severe levels of disease) have higher correlations amongst themselves (Bond, 1994). This means that the risk items may potentially belong to the same broad domain with the other items of CORE-OM, but load on different components due to their higher level of ‘difficulty’. Moreover, the finding of most positively worded items loading on the same component may be attributable to the common way of phrasing these items rather than their belonging to a domain that expresses a distinct attribute.

The advisory group considered the findings of PCA together with the issues raised above regarding the possibility of parameters other than the dimensionality of CORE-OM affecting the results of PCA. It was therefore agreed to attempt Rasch analysis on the whole CORE-OM, and not to carry out separate Rasch analyses on each domain identified in PCA, which is the approach that has been previously reported in the literature for the derivation of PBMs from existing CSMs (Young et al., 2009 & 2011). Such a decision was dictated by the lack of an explicit multidimensional structure of the CORE-OM, as PCA showed that the majority of CORE-OM items formed a large item pool belonging to 2 domains only, despite the multidimensional conceptual structure of the measure. Moreover, the results were not entirely consistent in the analyses performed across the total sample [N1500] and the two random sub-samples [N750a] and [N750b], so that underlying domains were not clearly and undoubtedly defined. Finally, and more importantly, PCA identified strong correlations between the different domains of the instrument. If such correlations were not taken into consideration during the development of the new health state classification, there was the danger of formation of implausible health states if standard techniques used for the generation of health states, such as orthogonal block designs, were applied. This issue was raised in Chapter 4 (section 4.1), and will be revisited in Chapter 6 (section 6.2.1), where a new approach for the identification of plausible health states, based on Rasch analysis, is described.

In conclusion, Rasch analysis was decided to be undertaken on the whole instrument at the next step of the process, in an attempt to exclude unsuitable items and ultimately develop a unidimensional scale that fits into the Rasch model.

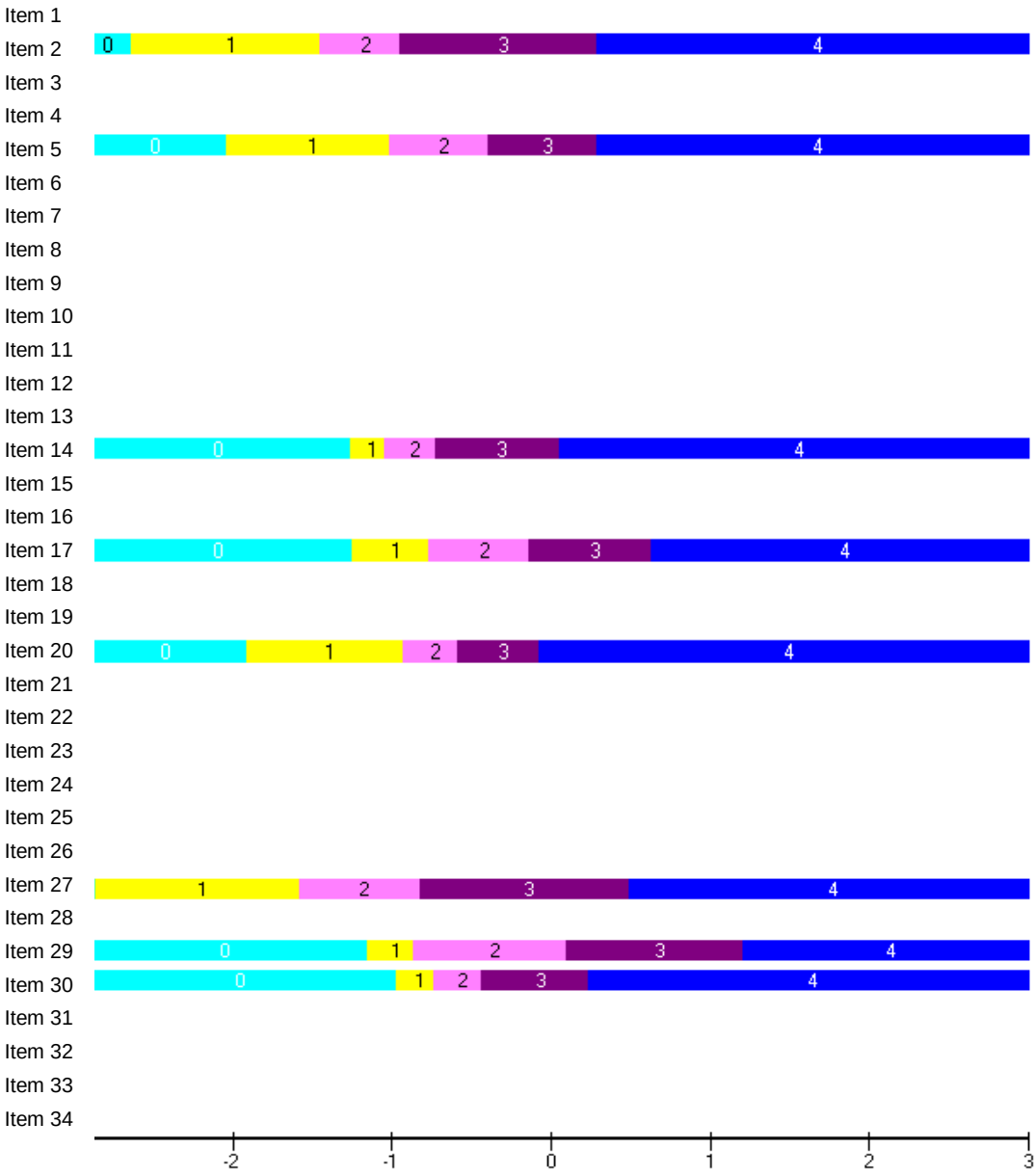
5.4 Results of Step 2: reduction of items and response levels from the CORE-OM

5.4.1 Rasch analysis

Investigation of threshold ordering and item rescoring

The first stage of Rasch analysis undertaken in [N400a] was to explore whether the CORE-OM items had ordered thresholds, i.e. whether respondents could discriminate between adjacent response levels. Items with disordered thresholds are demonstrated on the item threshold map, an output of RUMM2020 that provides the most likely responses to each item *with ordered thresholds* that are expected for each respondent, depending on the respondents' level of symptom severity as determined by their location on the Rasch logit scale. For items with disordered thresholds it is not possible to predict the most likely responses since respondents cannot differentiate across adjacent response levels; therefore the most likely responses for these items are not depicted on the Rasch item threshold map. As shown in Figure 8, which presents the item threshold map of Rasch analysis in [N400a], the majority of the CORE-OM items (24 out of the 34) had disordered thresholds. Only 8 items had ordered thresholds and were therefore depicted on the map, meaning that respondents could perceive the differences between the 5-level response statements relating to each of these items.

Figure 8. Rasch item threshold map of initial analysis of the CORE-OM – [N400a]



Response levels 0-4 correspond to response statements as follows: 0: not at all; 1: only occasionally; 2: sometimes; 3: often; 4: most or all the time, with the exception of positively worded items in which response statements are reversed.

The ‘threshold structure’ of each item can be visualised in a category probability curve, which shows the probability of every respondent obtaining each response level depending on respondents’ symptom severity expressed by their location on the Rasch logit scale.

Figure 9 shows the category probability curve for item 27 (I have felt unhappy), which is an item with ordered thresholds. It can be seen that as the location across the Rasch logit scale increases, the probability of moving from response level 0 (not at all) to response level 4 (most or all the time) through intermediate levels 1 (only occasionally), 2 (sometimes) and 3 (often) also increases; each response level receives its ‘peak’ probability at some location across the Rasch logit scale, with this location increasing as the response level increases. When a response level acquires its ‘peak’ probability, all other response levels have lower probabilities of being chosen by respondents, meaning respondents at this location are more likely to select the response level in question and less likely to select any other response level. In contrast, Figure 10 shows the category probability curve for item 22 (I have threatened or intimidated another person), which was found to have disordered thresholds. In this case the probability of obtaining consecutive levels of response does not increase with increasing person location; rather, persons seem to be able to distinguish only between response levels 0 (not at all) and 4 (most or all the time), with levels 1, 2 and 3 never becoming the most likely responses at any location across the scale. The category probability curves for all CORE-OM items as produced by initial Rasch analysis on [N400a] are illustrated in Appendix 6.

Figure 9. Example of category probability curve for an item with ordered thresholds [item 27 – I have felt unhappy]

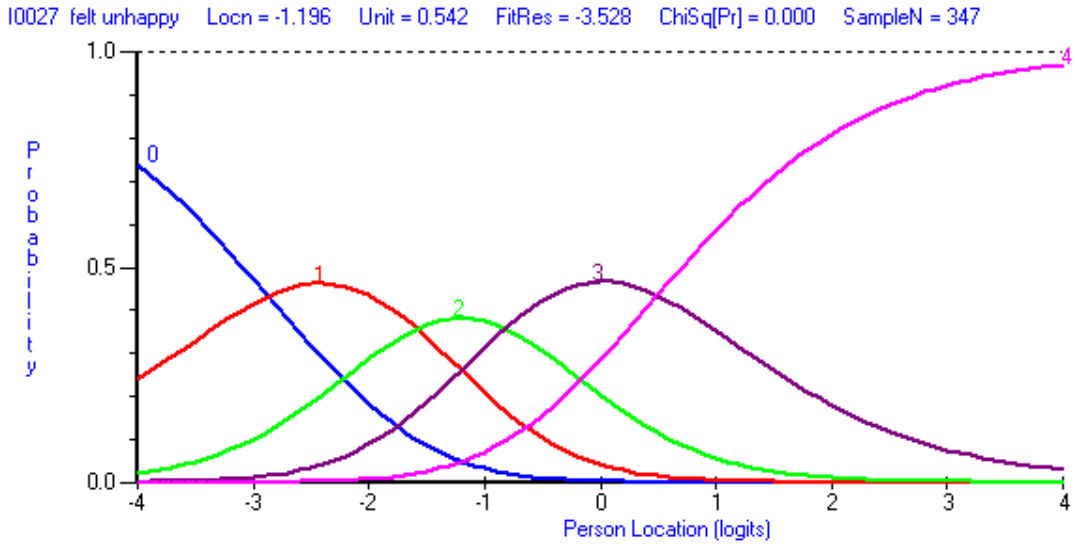
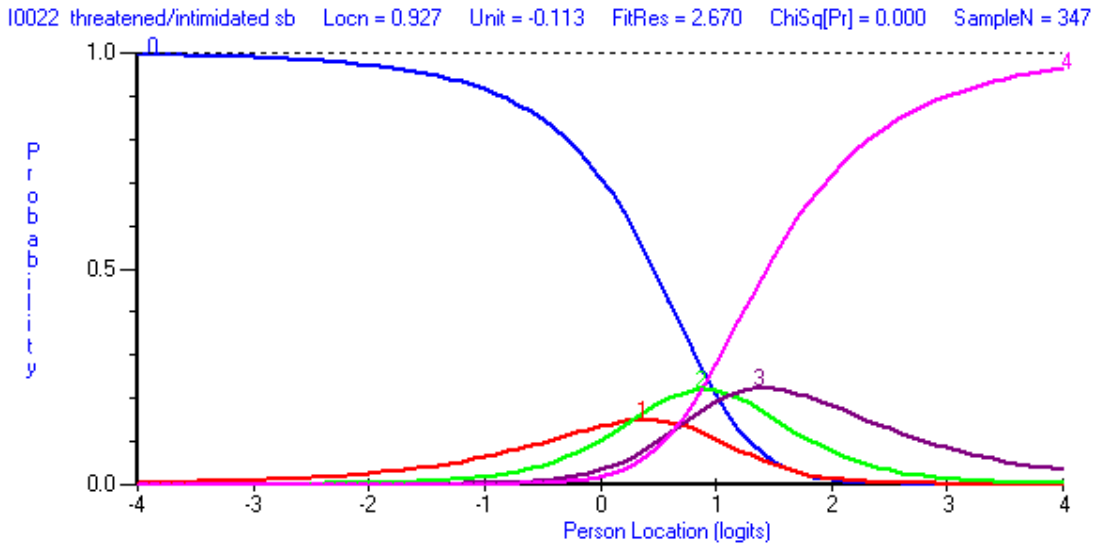


Figure 10. Example of category probability curve for an item with disordered thresholds [item 22 – I have threatened or intimidated another person]



The next stage before attempting to do any further Rasch analysis was to order the thresholds of all items. As described in section 4.2.2 of Chapter 4, at this stage the criteria for collapsing response levels in order to create new response categories were:

- visual inspection of category probability curves for each item: response levels with substantial overlapping in the areas under the curve were candidates for merging
- examination of category response proportions: response levels selected by very few respondents were generally candidates for merging with adjacent response levels
- minimum loss of information - thus the minimum possible merging of response categories was attempted
- clinically meaningful response statements.

Table 16 shows the category response proportions for each item, together with its status of threshold ordering. Items in the table have been ordered from the easiest to the most difficult one, according to their average location on the Rasch logit scale. It can be seen that, overall, easier items tend to have ordered thresholds. As difficulty increases, practically all items have disordered thresholds. Another point for observation is that difficult items (at the bottom of the table), especially risk items (which are the most difficult ones as they capture high symptom severity), have very low response proportions in their higher response levels. In contrast, easier items (on top of the table) tend to have low response proportions in their lower response levels (notably for response level 0 – “not at all”), although this finding does not seem to affect threshold ordering.

Table 16. Category response proportions for the 34 CORE-OM items before threshold ordering was attempted – [N400a]

Item	Location	Threshold	Proportion of responses in each response category				
			0	1	2	3	4
27	-1.196	Ordered	**0.03**	0.10	0.23	0.37	0.28
2	-1.188	Ordered	**0.02**	**0.09**	0.20	0.38	0.31
20	-0.877	Ordered	**0.06**	0.13	0.20	0.27	0.33
5	-0.790	Ordered	**0.05**	0.15	0.25	0.30	0.26
14	-0.745	Ordered	**0.08**	0.11	0.19	0.30	0.32
18	-0.606	*Disordered*	0.11	**0.08**	0.23	0.25	0.33
32	-0.573	*Disordered*	0.10	0.10	0.33	0.21	0.26
30	-0.478	Ordered	0.11	0.15	0.22	0.28	0.24
31	-0.468	*Disordered*	**0.09**	0.13	0.29	0.29	0.20
4	-0.464	*Disordered*	**0.09**	**0.08**	0.40	0.26	0.17
17	-0.379	Ordered	0.12	0.17	0.26	0.26	0.19
13	-0.347	*Disordered*	0.13	0.12	0.29	0.28	0.18
28	-0.316	*Disordered*	0.17	0.10	0.26	0.23	0.23
12	-0.232	*Disordered*	0.11	0.14	0.38	0.25	0.11
7	-0.212	*Disordered*	0.14	0.11	0.38	0.22	0.14
29	-0.177	Ordered	0.12	0.18	0.32	0.26	0.11
1	-0.142	*Disordered*	0.16	0.14	0.33	0.23	0.13
23	-0.141	*Disordered*	0.19	0.17	0.25	0.23	0.16
11	-0.111	*Disordered*	0.19	0.16	0.28	0.21	0.15
8	-0.067	*Disordered*	0.24	0.19	0.23	0.16	0.17
3	0.014	*Disordered*	0.23	0.12	0.28	0.25	0.12
25	0.056	*Disordered*	0.21	0.16	0.28	0.24	0.10
10	0.148	*Disordered*	0.22	0.21	0.34	0.14	**0.08**
15	0.196	*Disordered*	0.26	0.20	0.26	0.19	**0.09**
19	0.282	*Disordered*	0.34	0.26	0.24	**0.08**	**0.08**
21	0.349	*Disordered*	0.26	0.22	0.30	0.16	**0.06**
33	0.448	*Disordered*	0.42	0.15	0.27	**0.09**	**0.08**
26	0.509	*Disordered*	0.43	0.19	0.18	0.14	**0.06**
24	0.628	*Disordered*	0.53	0.20	0.10	0.10	**0.07**
22	0.927	*Disordered*	0.74	0.11	**0.08**	**0.03**	**0.03**
9	1.344	*Disordered*	0.69	0.14	0.11	**0.04**	**0.01**
16	1.464	*Disordered*	0.83	**0.09**	**0.06**	**0.01**	**0.01**
6	1.544	*Disordered*	0.88	**0.07**	**0.04**	**0.01**	**0.01**
34	1.598	*Disordered*	0.84	**0.09**	**0.05**	**0.01**	**0.01**

Response levels 0-4 correspond to response statements as follows: 0: not at all; 1: only occasionally; 2: sometimes; 3: often; 4: most or all the time, with the exception of positively worded items in which response statements are reversed. Proportions <0.10 are shown in bold.

Various ways of merging response levels and rescoring items with disordered thresholds were attempted using the set criteria, until all items demonstrated ordered thresholds in the item threshold map of Rasch analysis. Since no common re-scoring to all 34 CORE-OM items was possible to apply at this stage, the partial-credit Rasch model was selected for analysis of the data. Figure 11 shows the new response levels of CORE-OM items following merging of response statements, where required, so as to achieve threshold ordering. In the case of some items, for example item 34 (I have hurt myself physically or taken risks with my health), more than two adjacent response levels needed to be merged in order to acquire ordered thresholds.

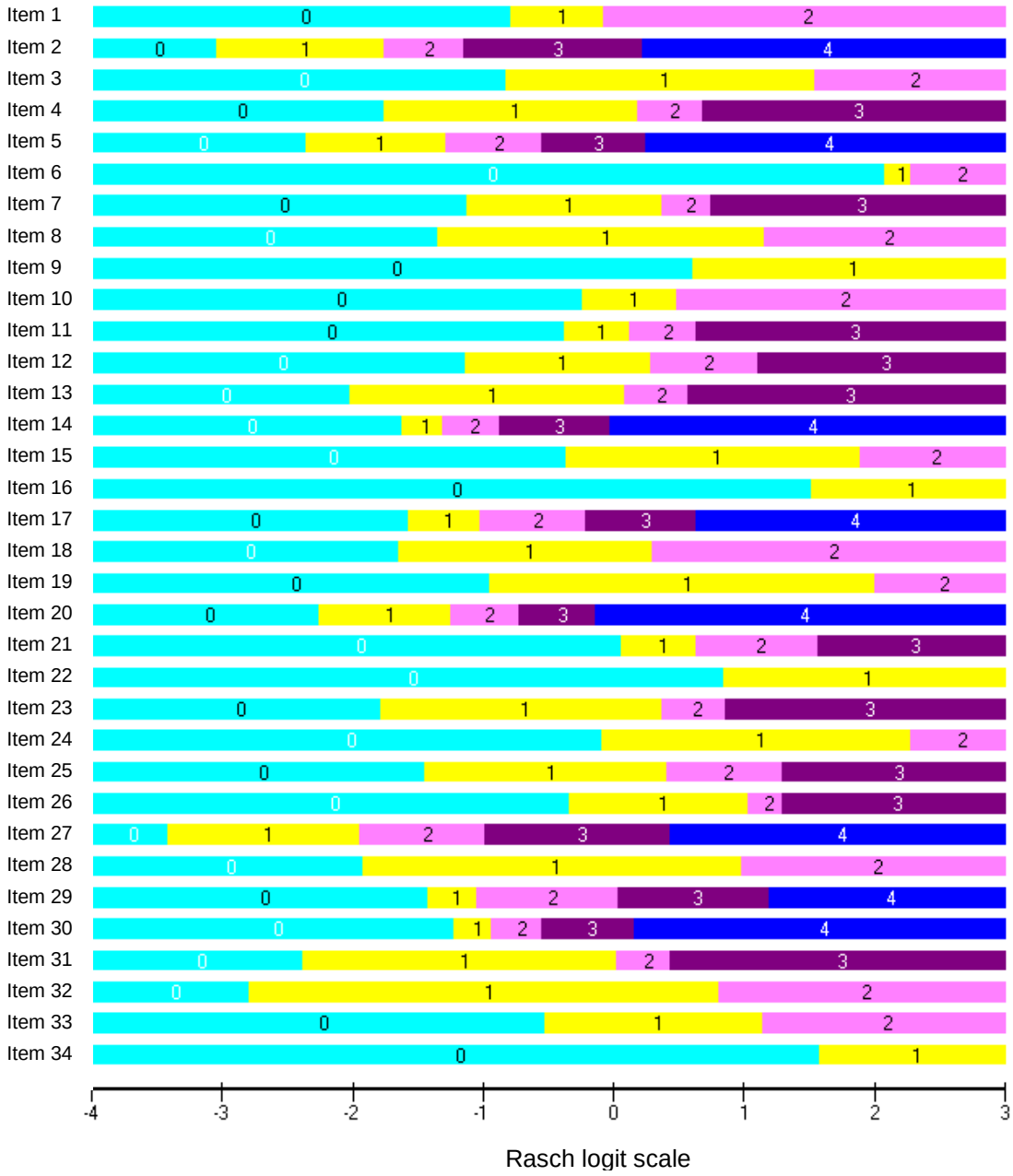
Figure 12 shows the Rasch item threshold map following threshold ordering of all CORE-OM items. The category probability curves of the 34 CORE-OM items following merging of response levels and threshold ordering are provided in Appendix 7.

Figure 11. New response levels of the CORE-OM items following merging of original response levels and subsequent threshold ordering – [N400a]

ITEM	0	1	2	3	4
1. [terribly alone and isolated]					
2. [tense, anxious or nervous]					
3. [somebody to turn to for support]					
4. [felt ok about myself]					
5. [totally lacking in energy and enthusiasm]					
6. [physically violent to others]					
7. [able to cope when things go wrong]					
8. [aches, pains, physical problems]					
9. [thought of hurting myself]					
10. [talking to people has felt too much]					
11. [tension/anxiety prevented doing things]					
12. [happy with the things I've done]					
13. [disturbed by unwanted thoughts/feelings]					
14. [felt like crying]					
15. [felt panic or terror]					
16. [made plans to end my life]					
17. [overwhelmed by my problems]					
18. [difficulty of getting to sleep/staying asleep]					
19. [felt warmth or affection for someone]					
20. [problems impossible to put to one side]					
21. [able to do most things I needed to]					
22. [threatened or intimidated another person]					
23. [felt despairing or hopeless]					
24. [thought it would be better if I were dead]					
25. [felt criticised by other people]					
26. [thought I have no friends]					
27. [felt unhappy]					
28. [unwanted images/memories distressing]					
29. [irritable when with other people]					
30. [I am to blame for problems & difficulties]					
31. [felt optimistic about my future]					
32. [achieved the things I wanted to]					
33. [felt humiliated or shamed by other people]					
34. [hurt myself physically/risks with health]					

Response levels 0-4 correspond to response statements as follows: 0: not at all; 1: only occasionally; 2: sometimes; 3: often; 4: most or all the time, with the exception of positively worded items in which response statements are reversed.

Figure 12. Rasch item threshold map of the CORE-OM after item rescoring (leading to ordered thresholds for all items) – [N400a]



Response levels 0-4 correspond to different response statements, depending on the merging of response levels – this map needs to be interpreted in connection with

Analysis after threshold ordering

Rasch analysis was undertaken on the CORE-OM following threshold ordering. According to the summary statistics shown in Table 17, the CORE-OM was somewhat 'difficult' for the study sample: its mean location on the Rasch logit scale was zero (set by default), while the mean location of the respondents was lower, at -0.328. The fit residual of the standard deviation of the mean location of items was high, at 2.283, implying some misfit in the model, while the respective parameter of persons equalled 1.496. The total chi-square probability for the model's degrees of freedom was <0.001, indicating that the CORE-OM did not fit in the Rasch model. The PSI, which expresses the reliability of a measure, was estimated at 0.93, indicating that the measure had excellent ability to discriminate amongst different respondent groups.

Table 17. Rasch analysis of CORE-OM data following threshold ordering: item-person and item-trait interaction summary statistics– [N400a]

	Items		Persons	
	Location	**Fit residual**	**Location**	**Fit residual**
Mean	0.000	0.098	-0.328	-0.207
Standard deviation	0.865	2.283	0.964	1.496

Item-trait interaction		**Reliability indices**	
Total Item chi-square	460.314	**PSI**	0.93104
Total degrees of freedom	170.000	**Cronbach Alpha**	N/A (missing data)
Total chi-square probability	0.00000	**Power of test-of-fit**	Excellent

Individual item statistics and indications for DIF are shown in Table 18. It can be seen that a number of items (items 1, 3, 8, 17, 19, 23, 27, 30 and 31) did not fit in the Rasch model as they demonstrated fit residuals beyond ± 2.5. Moreover, a number of items (items 3, 8, 9, 19, 23, 24, 27, and 30) showed significant chi-squared probabilities following Bonferroni adjustment, indicating their misfit in the Rasch model. Five items showed DIF: item 14 was characterised by uniform DIF by gender, while items 6, 8, 24 and 29 were

characterised by uniform DIF caused by age. In terms of location, items belonging to the conceptual domains of ‘symptoms’ and ‘subjective well-being’ tended to be easier than items belonging to ‘functioning’. Risk items appeared to be the most difficult ones. These results indicate that people with mild common mental health problems first experience a reduction in their subjective well-being, along with symptoms relating to emotional distress. As symptom severity increases, people experience problems relating to their functioning, both general and within their relationships. At the most severe level of common mental health problems, people seem to take risks with themselves and with other people. This finding on the relative ‘difficulty’ of the conceptual domains corroborates the suggested 3-phase model of psychotherapy outcome that entails projective improvement of subjective well-being, which is prerequisite for the (subsequent) reduction in emotional distress, which, in turn, is necessary before improvement in functioning is achieved (Howard et al., 1993). It also supports the concern that the PCA may ascribe dimensionality based on item difficulty.

Table 18. Rasch analysis of CORE-OM data following threshold ordering: individual item statistics and indications of differential item functioning – [N400a]

Item	CD	Location	Residual	Chi-square	P-value	DIF
1. [terribly alone and isolated]	FC	-0.430	**-2.904**	14.998	0.010	No
2. [tense, anxious or nervous]	SA	-1.430	-0.616	6.866	0.231	No
3. [somebody to turn to for support]	FC	0.364	**3.134**	**20.166**	**0.001**	No
4. [felt ok about myself]	SW	-0.296	-1.278	11.010	0.051	No
5. [totally lacking in energy and enthusiasm]	SD	-0.986	1.630	9.731	0.083	No
6. [physically violent to others]	RO	2.182	-0.394	13.651	0.018	**Yes**
7. [able to cope when things go wrong]	FG	0.003	0.619	4.722	0.451	No
8. [aches, pains, physical problems]	SP	-0.092	**3.547**	**23.140**	**0.000**	**Yes**
9. [thought of hurting myself]	RS	0.612	-1.516	**21.974**	**0.001**	No
10. [talking to people has felt too much]	FS	0.125	-0.296	0.794	0.977	No
11. [tension/anxiety prevented doing things]	SA	0.127	0.045	1.383	0.926	No
12. [happy with the things I've done]	FG	0.086	0.177	1.706	0.888	No
13. [disturbed by unwanted thoughts/feelings]	ST	-0.452	1.416	11.120	0.049	No
14. [felt like crying]	SW	-0.958	-1.024	1.940	0.857	**Yes**
15. [felt panic or terror]	SA	0.760	-0.083	3.304	0.653	No
16. [made plans to end my life]	RS	1.520	-1.168	14.745	0.012	No
17. [overwhelmed by my problems]	SW	-0.542	**-2.647**	13.309	0.021	No
18. [difficulty of getting to sleep/staying asleep]	SP	-0.669	0.637	5.091	0.405	No
19. [felt warmth or affection for someone]	FC	0.530	**4.642**	**53.685**	**0.000**	No
20. [problems impossible to put to one side]	SA	-1.090	0.276	1.258	0.999	No
21. [able to do most things I needed to]	FG	0.753	-0.591	10.146	0.071	No
22. [threatened or intimidated another person]	RO	0.847	1.259	8.202	0.145	No
23. [felt despairing or hopeless]	SD	-0.186	**-4.230**	**39.698**	**0.000**	No
24. [thought it would be better if I were dead]	RS	1.098	-1.926	**20.214**	**0.001**	**Yes**
25. [felt criticised by other people]	FS	0.087	-0.445	5.269	0.384	No
26. [thought I have no friends]	FC	0.660	-0.238	5.567	0.350	No
27. [felt unhappy]	SD	-1.476	**-4.086**	**29.228**	**0.000**	No
28. [unwanted images/memories distressing]	ST	-0.472	-0.582	13.432	0.020	No
29. [irritable when with other people]	FS	-0.307	2.491	9.778	0.082	**Yes**
30. [I am to blame for problems & difficulties]	SD	-0.636	**6.074**	**44.289**	**0.000**	No
31. [felt optimistic about my future]	SW	-0.640	**3.613**	19.361	0.002	No
32. [achieved the things I wanted to]	FG	-0.987	0.158	4.119	0.532	No
33. [felt humiliated or shamed by other people]	FS	0.308	-1.724	12.224	0.032	No
34. [hurt myself physically/risks with health]	RS	1.585	-0.643	4.195	0.522	No

Residuals ≥ $\left|2.5\right|$ are considered high; chi-squared probabilities have been assessed using Bonferroni adjustment. DIF was determined by chi-square statistics following Bonferroni adjustment. All statistics showing item misfit in the Rasch model are illustrated in bold. CD = conceptual domain; FC = functioning-close relationships; FG = functioning-general; FS = functioning-social relationships; SA = symptoms-anxiety; SD = symptoms-depression; SP = symptoms-physical; ST = symptoms-trauma; SW = subjective well-being; RO = risk/harm-to-others; RS = risk/harm-to-self.

5.4.2 Classical psychometric tests

The results of classical psychometric tests regarding responsiveness, floor or ceiling effects, correlation with the total CORE-OM score and percentage of missing data are shown in Table 19, while the full distribution of responses (indicating floor or ceiling effects) is shown in Table 16. Risk items had rather low responsiveness (<0.50), which, however, improved substantially once respondents with a baseline response of 'not at all' were removed from analysis. This finding was not unexpected, given that risk items are the most difficult ones, and therefore a high proportion of respondents were at the lowest response level at baseline with no scope for improvement. Item 19 had also low responsiveness (SRM 0.33), which did not improve much after exclusion of respondents with a zero response level at baseline. A number of items demonstrated floor effects (proportion of the respondents at the lower response level ≥ 0.30). These included all risk items, a finding that was anticipated given the difficulty of these items, which meant that a large proportion of participants gave a negative response ('not at all') at baseline. Another 3 items (19, 26 and 33) that were among the most difficult according to their mean location also showed floor effects. On the other hand, items 2, 14, 18 and 20 demonstrated ceiling effects (proportion of responders at the highest response level ≥ 0.30). Regarding correlation with the total CORE-OM score, items 34, 19, 8 and the two risk/harm-to-others items 6 and 22 showed rather low correlation as expressed by the Spearman's ρ value. The percentage of missing data was low for all items, with only item 19 showing somewhat higher percentage compared with the other CORE-OM items (2.4% vs. ≤ 1.5%).

Table 19. Results of standard psychometric tests on CORE-OM items: responsiveness, floor and ceiling effects, correlation with total CORE-OM score and percentage of missing data – [N400a]

Item	SRM (sub-analysis)	Pr of response at level 0 / 4*	Spearman's ρ value	Missing data
1. [terribly alone and isolated]	0.99 (1.31)	0.16 / 0.13	0.71	0.4%
2. [tense, anxious or nervous]	1.18 (1.25)	0.02 / **0.31**	0.60	0.3%
3. [somebody to turn to for support]	0.65 (0.97)	0.23 / 0.12	0.42	0.7%
4. [felt ok about myself]	1.00 (1.19)	0.09 / 0.17	0.65	0.6%
5. [totally lacking in energy and enthusiasm]	0.96 (1.13)	0.05 / 0.26	0.59	0.4%
6. [physically violent to others]	**0.24** (1.54)	**0.88** / 0.01	**0.28**	0.5%
7. [able to cope when things go wrong]	0.78 (1.05)	0.14 / 0.14	0.59	0.6%
8. [aches, pains, physical problems]	0.61 (0.92)	0.24 / 0.17	**0.28**	0.7%
9. [thought of hurting myself]	**0.46** (1.58)	**0.69** / 0.01	0.53	0.4%
10. [talking to people has felt too much]	0.81 (1.13)	0.22 / 0.08	0.55	0.7%
11. [tension/anxiety prevented doing things]	0.89 (1.23)	0.19 / 0.15	0.64	0.8%
12. [happy with the things I've done]	0.85 (1.11)	0.11 / 0.11	0.62	0.8%
13. [disturbed by unwanted thoughts/feelings]	0.95 (1.23)	0.13 / 0.18	0.56	0.5%
14. [felt like crying]	1.19 (1.40)	0.08 / **0.32**	0.63	0.3%
15. [felt panic or terror]	0.84 (1.36)	0.26 / 0.09	0.58	0.4%
16. [made plans to end my life]	**0.29** (1.54)	**0.83** / 0.01	0.44	**1.0%**
17. [overwhelmed by my problems]	1.09 (1.32)	0.12 / 0.19	0.74	**1.0%**
18. [difficulty of getting to sleep/staying asleep]	0.93 (1.09)	0.11 / **0.33**	0.52	0.6%
19. [felt warmth or affection for someone]	**0.33** (0.66)	**0.34** / 0.08	**0.30**	**2.4%**
20. [problems impossible to put to one side]	1.04 (1.16)	0.06 / **0.33**	0.63	0.9%
21. [able to do most things I needed to]	0.69 (1.05)	0.26 / 0.06	0.57	0.8%
22. [threatened or intimidated another person]	**0.32** (1.15)	**0.74** / 0.03	**0.27**	**1.0%**
23. [felt despairing or hopeless]	1.09 (1.43)	0.19 / 0.16	0.79	0.8%
24. [thought it would be better if I were dead]	0.58 (1.39)	**0.53** / 0.07	0.65	0.7%
25. [felt criticised by other people]	0.70 (1.07)	0.21 / 0.10	0.56	0.8%
26. [thought I have no friends]	0.65 (1.25)	**0.43** / 0.06	0.60	0.9%
27. [felt unhappy]	1.26 (1.32)	0.03 / 0.28	0.73	0.5%
28. [unwanted images/memories distressing]	0.89 (1.17)	0.17 / 0.23	0.58	0.6%
29. [irritable when with other people]	0.86 (1.07)	0.12 / 0.11	0.55	0.9%
30. [I am to blame for problems & difficulties]	0.80 (0.99)	0.11 / 0.24	0.53	0.5%
31. [felt optimistic about my future]	0.81 (0.96)	0.09 / 0.20	0.47	**1.0%**
32. [achieved the things I wanted to]	0.86 (1.07)	0.10 / 0.26	0.59	**1.5%**
33. [felt humiliated or shamed by other people]	0.61 (1.31)	**0.42** / 0.08	0.56	**1.1%**
34. [hurt myself physically/risks with health]	**0.27** (1.25)	**0.84** / 0.01	**0.35**	0.9%

SRM = standardised response mean; in parenthesis results for each item after including only respondents who had a baseline value of at least 1 in the particular item (or at maximum 3 in positively worded items). Pr = proportion of respondents; Spearman's ρ value expresses correlation with total CORE-OM score. In bold: SRM values <0.50; proportion of respondents at level 0 or 4 ≥ 30%; Spearman's ρ values < 0.40; and % of missing data ≥ 1.0%
*levels have been reversed for positively worded items.

5.4.3 Interpretation of the findings – advisory group's opinion on candidate items for exclusion from the new health state classification

The thesis advisory group considered the findings of Rasch analysis on the 34 CORE-OM items following threshold ordering, the results of the standard psychometric tests and other issues on the appropriateness and relevance of the items for inclusion in a PBM, in order to identify candidate items for exclusion from the new measure. The following views were expressed:

The two risk/harm-to-others items 6 (I have been physically violent to others) and 22 (I have threatened or intimidated another person) should be prioritised for deletion. Both items had very low correlation with the total CORE-OM score and demonstrated low responsiveness to treatment and floor effects (although these findings were partially justified by the high difficulty of these items). Moreover, item 6 showed DIF. Most importantly, the advisory group expressed the view that these items were not relevant to a PBM, as they expressed external behaviour affecting society rather than people's perceptions of their own HRQoL.

Item 34 (I have hurt myself physically or taken risks with my health) was characterised by low responsiveness, low correlation with the total CORE-OM score and floor effects. Moreover, the advisory group judged its wording to be ambiguous. Therefore, this item was also prioritised for exclusion.

Item 8 (I have been troubled by aches, pains, physical problems) was also decided to be excluded due to significant item misfit into the Rasch model, which had been anticipated, since the item expressed physical symptoms and therefore clearly belonged to a different dimension from items measuring, in their majority, emotional symptoms. In addition, item 8 demonstrated DIF and, not surprisingly, low correlation with the total CORE-OM score. 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, at a later stage, with the final (unidimensional) product of Rasch analysis.

Other items that demonstrated misfit to the model as indicated by either their fit residuals or their chi-squared probability (that is, items 1, 3, 9, 17, 19, 23, 24, 27, 30 and 31) were considered for exclusion from the final measure. Items 3, 19 and 31 also showed relatively low correlation with the total CORE-OM score. Moreover, items 19, 26 and 33 demonstrated floor effects, and item 19 had the highest percentage of missing data.

Items 14 and 29 were potential candidates for exclusion as they demonstrated DIF in the initial analysis. DIF was considered non-acceptable in a PBM that should capture HRQoL of all respondents in a similar way, without systematically discriminating according to demographic characteristics.

It should be noted that the initial misfit of items into the Rasch model was only an indication for exclusion from the final instrument, and did not determine exclusion at this stage. This was decided because item fit in the Rasch model depends to a large extent on the behaviour of the other items comprising the scale, and therefore exclusion of other items might alter item fit statistics at later stages of analysis. Therefore, items were excluded one at a time followed by Rasch analysis on the remaining items and subsequent testing of fit statistics. The order of exclusion of items was agreed with the advisory group, based on the group's considerations as previously described. This process was repeated until all remaining items fit in the Rasch model.

5.4.4 Exclusion of unsuitable CORE-OM items - construction of a scale fitting the Rasch model

Items 6, 22, 34 and 8 were the first items that were excluded from further consideration, based on the findings of the analyses and the advisory group's views. Successive Rasch analyses led to the exclusion of items 3, 9, 19, 23, 24, 27, 30 and 31 that persistently (in the initial and all consecutive analyses) misfit into the Rasch model. Items 14 and 29 were excluded because they demonstrated persistently significant DIF. Items 5, 18 and 28, although did fit in the Rasch model in the initial analysis, showed high fit residuals ($\geq |2.5|$) at later stages and were eventually excluded from further consideration. On the other hand, items 1 and 17, which showed misfit in the model at initial stages

of analysis, appeared to fit in the model at later stages following deletion of other items, and were thus retained in the analysis. All item exclusions were agreed with the advisory group.

The 17 remaining items of the CORE-OM that fit in the Rasch model and their respective item fit statistics are presented in Table 20. The overall Rasch statistics of this 17-item measure are provided in Table 21. It can be seen that, according to the conceptual framework of CORE-OM, 9 of the 17 items reflect functioning; 5 items express symptoms; 2 items represent subjective well-being; and one item expresses risk/harm-to-self. Data on both tables indicate that the scale and all individual items fit the Rasch model (total chi-squared probability 0.17). No item was associated with DIF. The scale had an excellent ability to discriminate amongst different groups of respondents (PSI 0.90). The Rasch item threshold map of the 17-item scale illustrating the ordered thresholds of all items is shown in Figure 13.

Table 20. Results of Rasch analysis of the 17 CORE-OM items fitting the Rasch model: individual item statistics and indications of differential item functioning – [N400a]

Item	CD	Location	Residual	Chi-square	P-value	DIF
1. [terribly alone and isolated]	FC	-0.392	-1.393	8.602	0.126	No
2. [tense, anxious or nervous]	SA	-1.502	-0.326	3.543	0.617	No
4. [felt ok about myself]	SW	-0.270	0.053	1.902	0.863	No
7. [able to cope when things go wrong]	FG	0.068	0.631	7.073	0.215	No
10. [talking to people has felt too much]	FS	0.193	0.628	4.120	0.532	No
11. [tension/anxiety prevented doing things]	SA	0.218	0.008	6.176	0.289	No
12. [happy with the things I've done]	FG	0.147	0.925	1.141	0.950	No
13. [disturbed by unwanted thoughts/feelings]	ST	-0.405	2.379	11.933	0.036	No
15. [felt panic or terror]	SA	0.858	0.203	5.160	0.397	No
16. [made plans to end my life]	RS	1.579	-0.463	4.754	0.447	No
17. [overwhelmed by my problems]	SW	-0.525	-2.058	12.036	0.034	No
20. [problems impossible to put to one side]	SA	-1.109	0.275	3.177	0.673	No
21. [able to do most things I needed to]	FG	0.873	-0.454	5.300	0.380	No
25. [felt criticised by other people]	FS	0.140	0.996	3.811	0.577	No
26. [thought I have no friends]	FC	0.731	0.873	9.621	0.087	No
32. [achieved the things I wanted to]	FG	-0.987	0.918	1.139	0.951	No
33. [felt humiliated or shamed by other people]	FS	0.384	-0.897	7.631	0.178	No

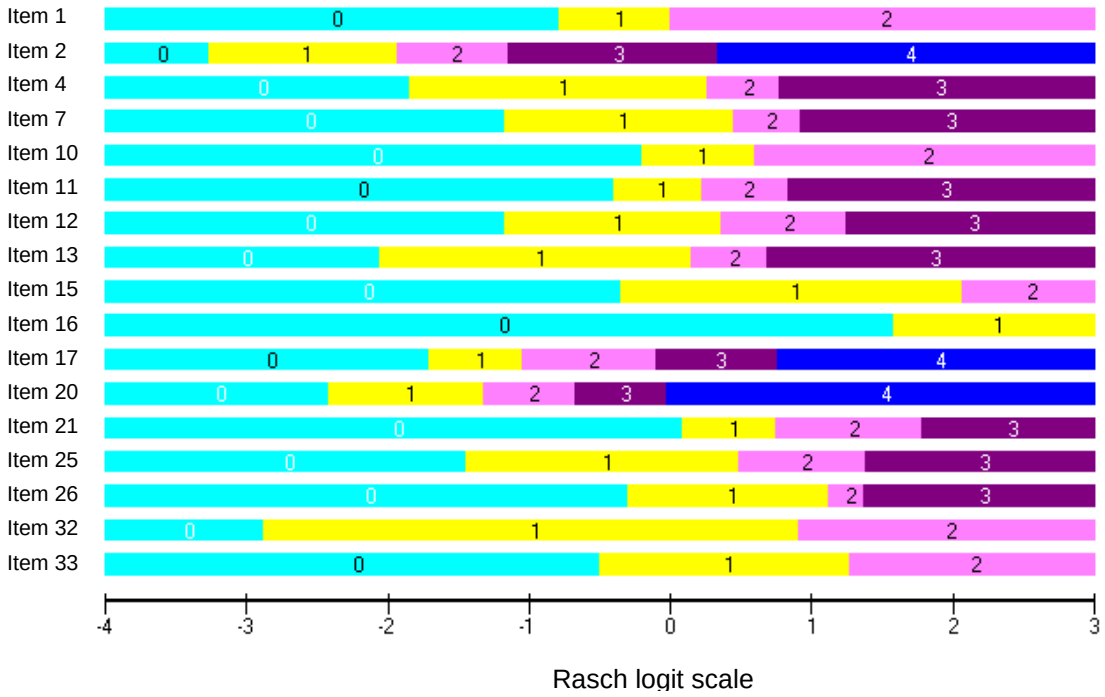
Residuals beyond ≥ |2.5| are considered high; chi-square probabilities have been assessed using Bonferroni adjustment. DIF was determined by chi-square statistics following Bonferroni adjustment. CD = conceptual domain; FC = functioning-close relationships; FG = functioning-general; FS = functioning-social relationships; RS = risk/harm-to-self; SA = symptoms-anxiety

Table 21. Rasch analysis of the 17 CORE-OM items fitting the Rasch model: item-person and item-trait interaction statistics – [N400a]

	Items		Persons	
	Location	Fit residual	Location	Fit residual
Mean	0.000	0.135	-0.304	-0.220
Standard deviation	0.788	1.037	1.143	1.223

Item-trait interaction		Reliability indices	
Total Item chi-square	97.117	**PSI**	0.90087
Total degrees of freedom	85	**Cronbach Alpha**	N/A (missing data)
Total chi-square probability	0.1737	**Power of test-of-fit**	Excellent

Figure 13. Rasch item threshold map of the 17-CORE-OM item scale fitting the Rasch model – [N400a]



5.5 Results of Step 3: selection of CORE-OM items for inclusion in the new health state classification

Although the 17-item scale derived from the CORE-OM fit the Rasch model, further reduction was required in order to construct a concise measure amenable to valuation. Further exclusion of items was thus undertaken, after testing different item combinations and applying the set criteria reported in section 4.2.3 of Chapter 4.

Coverage of CORE-OM domains

The advisory group expressed the view that the items in the final instrument should capture as many domains of the CORE-OM as possible, as represented in the measure's conceptual domains but also as indicated by PCA. This would ensure that the new PBM is able to tap a range of different aspects of HRQoL that are relevant to people with common mental health problems, as captured by the CORE-OM, with minimum loss of information. In

addition, the new measure should include at least one positively worded item to retain the ‘character’ of the original measure. However, it was acknowledged that, since the measure needed to fit in the Rasch model and fulfil a number of other criteria, representation of all domains might not be possible.

Of the 10 conceptual sub-domains of the CORE-OM, 3 were not represented in the 17-item scale developed in the previous step, as their items had been excluded in previous stages of analysis: ‘symptoms - depression’, ‘symptoms - physical’, and ‘risk/harm to others’. It has to be noted, though, that item 8 was initially removed with the intension to be combined with the unidimensional final output of Rasch analysis at a later stage.

The ‘subjective well-being’ domain (items 4 and 17 in the 17-item scale) was also deemed as an important aspect of the measure; however, it was recognised that this domain covered the overall perception of persons’ HRQoL rather than distinct symptoms / problems of people with mental disorders. Moreover, in the PCA undertaken in Step 1 of the process, items belonging to ‘subjective well-being’ loaded on the same components with items expressing symptoms (components 1 and 2). Indeed, this domain has been previously found to highly correlate with items in the overall ‘problems’ domain (Evans et al., 2002). It was therefore accepted that it was less crucial for this domain to be included in the final measure. Regarding the ‘symptoms - trauma’ sub-domain (item 13 in the 17-item scale), this was considered less relevant for inclusion in a HRQoL measure for people with common mental health problems. Importantly, attempts to include items of ‘subjective well-being’ and ‘symptoms - trauma’ in the final measure resulted in a scale not satisfying the final criterion of increase in item threshold locations with increasing difficulty of the items. Consequently, these two sub-domains were not represented in the final measure.

The advisory group concluded that the remaining conceptual sub-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

relationships' (items 10, 25, 33 in the 17-item scale) and 'risk/harm to self' (item 16 in the 17-item scale) reflected conceptual major domains in people with common mental health problems and should be represented in the final construct.

Furthermore, it was agreed that the new measure should ideally include one item per domain as identified in PCA undertaken on the CORE-OM. According to the results of this analysis, the vast majority of the items comprising the 17-item scale loaded on PCA components 1 [items 1, 2, 13, 17 and 20] and 2 [items 2, 4, 7, 11, 12, 15, 17, 21 and 32] with items 2 and 17 loading on both components. These two components included the majority of CORE-OM items, which belonged to various conceptual domains of CORE-OM. Three items [25, 26 and 33] loaded on component 3, which appeared to capture mostly items on functioning – relationships, and item 16 loaded on component 4, which contained all the risk-to-self items. Item 10 did not load on any of the components identified in PCA. On the other hand, the 5th significant component of PCA was the only one not represented in the 17-item scale.

Ensuring consistency in response levels across items of the new measure

Items were excluded one at a time and Rasch statistics as well as the PSI were constantly checked. In addition, during this process, a number of items were re-scored, while the impact of re-scoring on their threshold ordering as well as on the overall model and individual item fit was constantly checked. Rescoring of some items was attempted so that the final measure had homogeneous response levels across all its items (the items comprising the 17-item scale had different response levels). When item re-scoring aiming at consistency of response levels across all items of the final measure was not possible without negatively affecting overall model and individual item statistics, then the item was excluded from further analysis.

Constructing the emotional component of the new health state classification

Following exclusion of a number of items, various combinations of 5 items (of those included in the 17-item scale), corresponding to the 5 conceptual

domains considered crucial for representation in the new measure, were tested against the set criteria for this step, in order to construct the emotional component of the final health state classification. Testing of various item combinations resulted in a scale consisting of 5 items (1, 15, 16, 21 and 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 sub-domains, respectively, and included one positively worded item. Moreover, the 5 items represented 4 out of the 5 significant components that were identified in PCA, with component 2 being represented by 2 items, one of which was positively worded.

Model and individual item fit – reliability of the emotional component

The overall model statistics of the 5-item emotional component of the new measure are shown in Table 22. The scale demonstrated good model fit (chi-square probability 0.69). The measure appeared to be somehow ‘difficult’ for the study population, given the lower mean location of persons compared with the mean location of the items. However, this was attributed to the inclusion of one risk item (16), which was deemed necessary in order to capture more severe cases. The PSI index reached 0.66, which is somewhat lower than the 0.70 value that is generally considered acceptable for group comparison (Fisher, 1992). Nevertheless, the figure of 0.66 was deemed adequate for the purpose of the development of a new PBM by the advisory group, 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. All items fit in the model, as indicated by the statistics shown in Table 23; no DIF was observed in any of the items. The class interval structure, demonstrated in Table 24, showed a homogeneous allocation of respondents across class intervals.

Table 22. Rasch analysis of the final 5-item emotional component of the new measure: item-person and item-trait interaction statistics – [N400a]

	Items		Persons	
	Location	**Fit residual**	**Location**	**Fit residual**
Mean	0.000	0.072	-0.818	-0.254
Standard deviation	1.293	0.404	1.425	0.791

Item-trait interaction		**Reliability indices**	
Total Item chi-square	20.970	**PSI**	0.65929
Total degrees of freedom	25	**Cronbach Alpha**	N/A (missing data)
Total chi-square probability	0.6943	**Power of test-of-fit**	Good

Table 23. Results of Rasch analysis of the 5-item emotional component of the new measure: individual item statistics – [N400a]

Item	CD	Location	Residual	Chi-square	P-value
1. [terribly alone and isolated]	FC	-1.468	-0.099	2.044	0.843
15. [felt panic or terror]	SA	-0.881	-0.058	3.403	0.638
16. [made plans to end my life]	RS	1.801	-0.358	5.812	0.325
21. [able to do most things I needed to]	FG	0.702	0.717	6.520	0.259
33. [felt humiliated or shamed by other people]	FS	-0.154	0.156	3.191	0.671

Residuals ≥ $\left|2.5\right|$ are considered high; chi-square probabilities have been assessed using Bonferroni adjustment. CD = conceptual domain; FC = functioning-close relationships; FG = functioning-general; FS = functioning-social relationships; RS = risk/harm-to-self; SA = symptoms-anxiety

Table 24. Class interval distribution of the emotional component of the new measure – [N400a]

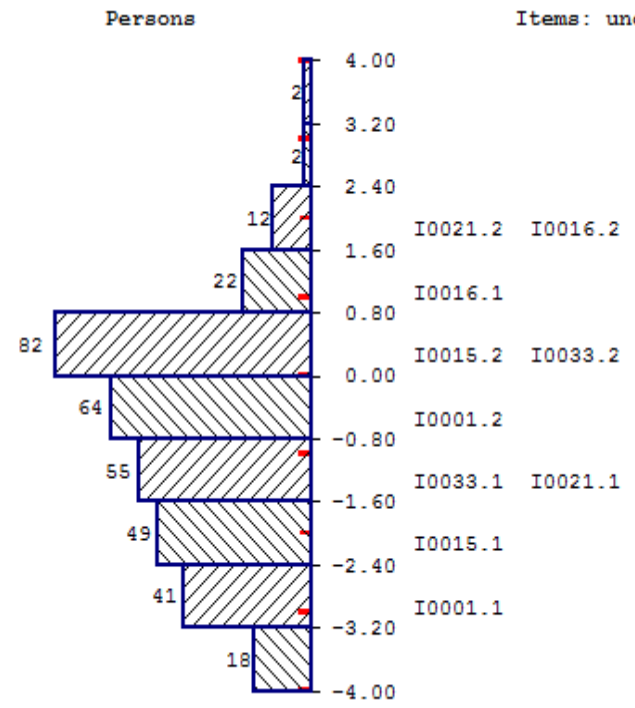
ITEM	CI1	CI2	CI3	CI4	CI5	CI6	CI7
1	42	46	56	63	46	37	35
15	41	46	56	64	46	37	35
16	42	46	56	63	47	37	35
21	41	46	54	63	46	37	35
33	42	46	56	62	46	37	35

CI: Class interval

Coverage of the full range of emotional symptom severity

The item map depicted in Figure 14 demonstrates that the instrument is well targeted to the study population as it is able to practically capture the full range of severity of emotional symptoms, with minimal floor or ceiling effects and good spread of items across the full range of respondents' scores.

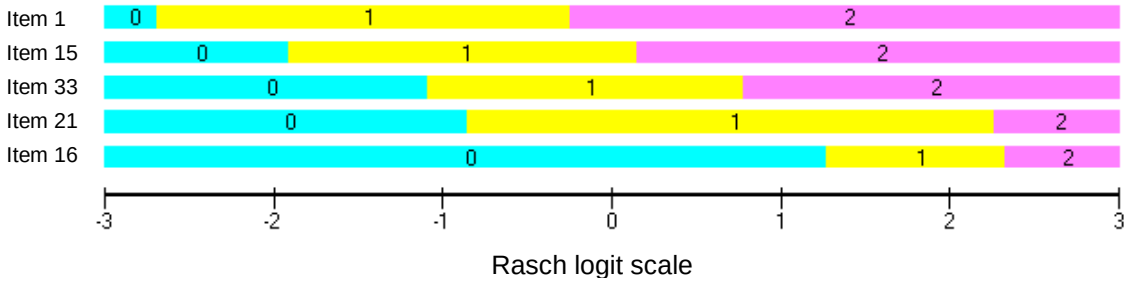
Figure 14. Item map of the emotional component of the new measure



Increase in item threshold locations with increasing difficulty of the item in the emotional component

The Rasch item threshold map of the emotional component of the new measure is shown in Figure 15. Items have been ordered from the easiest to the most difficult one according to their location on the Rasch model logit scale. 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.

Figure 15. Rasch item threshold map of the emotional component of the new health state classification system that was derived from the CORE-OM



Local item independence in the emotional component

PCA on the item fit residuals was undertaken to explore potential item correlations and therefore to test the local independence of the items. According to the Varimax Rotation loadings shown in Table 25, each item loaded highly on each of the 5 residual components identified, indicating that none is highly correlated with the others. This indication was confirmed in the residual correlation matrix, in Table 26, which showed low correlations in pair-wise comparisons between the 5 items.

Table 25. Principal Component Analysis on the item fit residuals: loadings of the 5 items of the emotional component of the new measure - [N400a]

Item	Component 1	Component 2	Component 3	Component 4	Component 5
1	0.20	**0.93**	-0.18	-0.13	-0.22
15	0.26	-0.23	-0.21	-0.11	**0.91**
16	0.13	-0.11	-0.14	**0.97**	-0.08
21	0.15	-0.17	**0.95**	-0.15	-0.18
33	**-0.91**	-0.22	-0.17	-0.16	-0.26

Varimax rotation; loadings ≥ │0.40│ are shown in bold

Table 26. Residual correlation matrix of the 5 items of the emotional component of the new measure – [N400a]

Item	1	15	16	21	33
1	1.00				
15	-0.31	1.00			
16	-0.16	-0.09	1.00		
21	-0.24	-0.26	-0.22	1.00	
33	-0.28	-0.37	-0.20	-0.19	1.00

None of the correlation coefficients between different items is ≥ │0.40│

Unidimensionality of the emotional component

The post-hoc test of unidimensionality (Smith, 2002) established the unidimensionality of the scale: items were divided between those positively and those negatively correlated with the first residual component, and subsequently, for each subset, two scores were estimated for each respondent. Independent t-tests between the two scores were undertaken for each respondent; 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 the new measure.

5.6 Results of Step 4: validation of the emotional component of the new health state classification

The emotional component of CORE-6D was validated by undertaking Rasch analysis on the random sample [N400b]. Validation was also achieved in an analysis adjusted for sample size (so as to avoid Type I error) on the whole initial sample [N1500]. Finally, the emotional component of CORE-6D was validated on the 'mixed' random sample [N1500v] that consisted of people with mental disorders presenting to either primary or secondary care, adjusted for sample size as well. In all 3 analyses the 5-item scale had satisfactory overall and item fit statistics and showed no DIF. Moreover, no DIF was observed for access site (primary or secondary care setting) in the analysis of [N1500v]. The post-hoc unidimensionality test verified the scale's unidimensionality in all 3 validation datasets; the item threshold map produced in the 3 validation analyses confirmed a smooth transition of responses from milder to more severe symptoms, the same with that demonstrated by the analysis on sample [N400a]. Results of the validation of the emotional component of CORE-6D are provided in Appendix 8.

5.7 Constructing a 2-dimensional health state classification: the development of CORE-6D

The 5-item emotional component derived from Rasch analysis was combined with the physical item 8, thus producing a new health classification tapping 5 emotional domains and one physical domain, named CORE-6D (Clinical Outcomes in Routine Evaluation – 6-dimensional health state classification). The response statements were slightly reworded in present tense, since statements referring to the present were deemed to be more appropriate for the respondents in a valuation survey to comprehend and value. Moreover, the response level 'not at all' of CORE-OM was replaced by 'never' in CORE-6D, as the latter was deemed to be more suitable for inclusion in a full statement expressing a HRQoL state.

The CORE-6D is a 6-item health descriptive system consisting of a 5-item unidimensional emotional component and a physical item. Each item has 3 response levels: ‘never’, ‘only occasionally or sometimes’ and ‘often, most or all the time’. Each response level gets an individual score (0-1-2). One item is positively worded and therefore its response levels are reversed. The total score of the emotional component is the sum of individual scores (ranging from 0 to 10), with higher scores indicating higher levels of emotional distress. This ordinal score can be converted into an interval score on the Rasch model logit scale, which is important for the modelling approach that was used to predict utility values for all states of the CORE-6D following the valuation survey, as it will be discussed in Chapter 6. The unidimensional emotional component combined with the physical item creates a 2-dimensional scale, tapping emotional and physical symptoms in people with common mental health problems. The system describes $3^6 = 729$ unique health states. The 6 CORE-OM items that formed the CORE-6D were slightly altered (mainly changed from past perfect to present tense) so as to construct response statements that are meaningful to respondents in a valuation survey. Other than that, it was intended that the response statements were kept as similar as possible (in terms of structure and phrasing) to the items of the original instrument, so that the new utility index can be unambiguously applied/ mapped on datasets containing the CORE-OM.

The CORE-6D health state descriptive system is shown in Table 27.

5.8 Discussion and conclusion

This chapter describes the development of CORE-6D, a 2-dimensional health state classification for common mental health problems that consists of an emotional component and a physical item. CORE-6D was derived from the CORE-OM using predominantly Rasch analysis. Due to the large dependence across the domains of CORE-OM, the methodology employed was different from that described in section 3.3.1 of Chapter 3, where Rasch analysis was used to derive health state classifications from CSMs with clear multidimensional structure. In the studies reviewed in that section, Rasch analysis was performed separately on each dimension of a multidimensional CSM, to identify the best performing item within each dimension; subsequently,

these items were combined in a multidimensional health state classification. In contrast, in the study carried out for this thesis, Rasch analysis was undertaken on the full measure (CORE-OM), aiming to identify, and eventually discard, items that differentiated from the largely unidimensional behaviour of the measure. Use of Rasch analysis in this case led to the development of the emotional component of CORE-6D and confirmed its unidimensional character.

Table 27. The CORE-6D health state classification

Emotional component		
1	I never feel terribly alone and isolated	0
	I feel terribly alone and isolated only occasionally or sometimes	1
	I feel terribly alone and isolated often, most or all the time	2
2	I never feel panic or terror	0
	I feel panic or terror only occasionally or sometimes	1
	I feel panic or terror often, most or all the time	2
3	I never feel humiliated or shamed by other people	0
	I feel humiliated or shamed by other people only occasionally or sometimes	1
	I feel humiliated or shamed by other people often, most or all the time	2
4	I am able to do most things I need to often, most or all the time	0
	I am able to do most things I need to only occasionally or sometimes	1
	I am not able to do the things I need to	2
5	I never make plans to end my life	0
	I make plans to end my life only occasionally or sometimes	1
	I make plans to end my life often, most or all the time	2
Physical health item		
6	I am never troubled by aches, pains or other physical problems	0
	I am troubled by aches, pains or other physical problems only occasionally or sometimes	1
	I am troubled by aches, pains or other physical problems often, most or all the time	2

CORE-6D was derived from the CORE-OM by applying Rasch analysis on data from a sample of people with common mental health problems presenting to NHS primary care counselling services in the UK. The new measure was validated on a large mixed sample of people with common mental health problems presenting to either primary or secondary care. Rasch analysis on this mixed sample illustrated that there was no DIF regarding the access site, thus confirming that the measure is applicable to (and does not differentiate between) people with common mental health problems treated in primary *and* secondary care settings. It should be noted, though, that primary care is currently the dominant service provider for people with common mental health problems in the UK, and the choice between primary and secondary care is more related to access issues determining the pathway into the service rather than to the patients' level of symptom severity. In any case, the 5-item unidimensional emotional component of CORE-6D is able to capture a broad range of severity of emotional symptoms in people with common mental health problems.

Further to the elimination of items and the development of a health state classification, the great advantage of the use of Rasch analysis in the case of the CORE-OM (and other measures with no clear multidimensional structure) is that it enables the identification of plausible health states amenable to valuation, thus preventing the generation of implausible health states that might occur following use of standard techniques (e.g. orthogonal block designs) usually employed for the generation of health states from multidimensional health state classifications. This issue is illustrated and further discussed in Chapter 6.

Chapter 6. Development of a preference-based index: valuation of CORE-6D

6.1 Introduction

This chapter reports on the methods and the results of the valuation survey that was undertaken in order to attach utility values to selected health states of the CORE-6D and describes the modelling techniques that were subsequently employed in order to develop an algorithm that links all the health states described by the new measure with appropriate utility values. This process led to the construction of a preference-based index that can be used for the estimation of QALYs in economic evaluation of interventions for common mental health problems. The analyses and findings presented in this chapter have also been reported in a publication by Mavranezouli and colleagues (2012).

6.2 Methods

The methods adopted for the valuation of CORE-6D, including the generation and selection of health states for use in the valuation survey and the modelling methods employed for the valuation of all health states described by the new measure, were dictated by the unidimensionality of the emotional component of CORE-6D, which did not allow use of conventional statistical techniques that have been previously undertaken for the development of other PBMs.

6.2.1 Generating plausible health states for the valuation survey using Rasch analysis

As already described in Chapter 1 (section 1.3.1), the method for the selection of health states for consideration in the valuation survey depends on the modelling approach that is used to predict utility values for all states of the health state classification following the valuation survey. In the composite modelling approach, which uses statistical modelling, the selection of health states for valuation can be made using a statistical design such as an orthogonal array or a balanced methodology. In the decomposed modelling approach, which is based on MAUT, every dimension of a measure is valued separately, followed by valuation of 'corner' multidimensional ('full') states,

which consist of one dimension at one extreme (usually the worst response level) and the rest dimensions at the other extreme (usually the best response level). As argued in Chapter 3 (section 3.3.1), both approaches require that dimensions be structurally independent from each other in order to create meaningful health states. Use of these conventional approaches for generating health states from a measure with high correlation between its items (such as a unidimensional measure or a measure with large unidimensional components) is not appropriate because it is likely to generate implausible health states that cannot be processed and valued by participants in a valuation survey.

CORE-6D consists of a 5-item emotional component and a physical item. The emotional component of CORE-6D is, by construction, unidimensional, meaning that its items are not independent from each other, resulting in some item response combinations being implausible; e.g. “I make plans to end my life often, most or all the time” and “I never feel terribly alone and isolated”. Therefore, conventional methods for generating health states were not appropriate in the case of CORE-6D. Instead, identification of plausible health states described by the emotional component of CORE-6D was based on the results of Rasch analysis and was achieved by a novel method developed for this thesis, named the ‘Rasch vignette approach’.

The Rasch vignette approach relies on the inspection of the Rasch item threshold map, an output of Rasch analysis that depicts the most likely item response combinations expected for *each* location across the Rasch model logit scale; this means that the map can help identify *the most likely* response combination at each level of emotional distress captured by the emotional component of CORE-6D, from the mildest to the most severe. These unique response combinations represent ‘emotional’ health states that have been observed in people with common mental health problems across the continuum of severity of emotional distress, and therefore they describe actual and, very importantly, *plausible* health states that are amenable to valuation. It should be clarified that the Rasch item threshold map allows identification of *the one most likely* (and thus plausible) health state at each location across the

continuous Rasch model scale; it does not depict *every* plausible health state described by a unidimensional scale. For each level of symptom severity there may be several other plausible health states (which have the same total ordinal scale score and Rasch model logit value) that are not depicted on the map, as they are less likely to be observed in the study population in comparison with the depicted state of that particular severity level.

Following identification of plausible health states in the Rasch item threshold map produced by Rasch analysis on the dataset [N400a], the 'emotional' health states of CORE-6D were combined with different response levels of the physical item, so as to produce 'full' CORE-6D health states, as described in section 6.3.1 later in this chapter. Given the way these states were generated, there were important implications for the design of the valuation survey, which are described below.

6.2.2 Valuation survey

A valuation survey using face-to-face interviews was carried out in South Yorkshire, aiming at determining public preferences for a number of health states derived from CORE-6D. Selected health states were valued using the TTO technique, which has been described in Chapter 1 (section 1.3.1). More specifically, the version of TTO developed by the MVH group was used, including the visual props designed by this group, i.e. a set of health state cards (including a 'full health' card and a 'death' card) and a double-sided time board, with one side used for states considered better than death and the other one for states rated as worse than death, that contains a sliding scale that can move across the board (in a range between 0 and 10 years) to show the number of years to be spent in each alternative option assessed (Dolan et al., 1996; Gudex, 1994).

According to the MVH protocol, respondents were first asked whether they preferred to live in a specified health state h_i for t = 10 years after which they died, or to die immediately. This question determined whether respondents valued the health state as better, worse, or equal to being dead. For health states considered better than death, respondents were asked to choose either life in the health state h_i for 10 years followed by death or life in full health for x

years where $x < 10$. The number of x years in full health was varied by units of one year, starting from $x = 5$, until the point where the respondent was indifferent or switched preferences between the two alternatives. The utility value given to the state h_i was $x/10$. For health states considered worse than being dead, respondents were asked to choose either life in the health state h_i for y years followed by full health for x years after which they would die (with $y + x = 10$), or immediate death. Years in full health (x) were varied by one year, starting from $x = 5$, by concurrently varying years in the health state (y) so that $y + x$ always equalled 10, until the point where respondents were indifferent or switched preferences between the two options. Valuations in the case of states considered worse than death were estimated using the formula $-x/10$, following the same process with that reported at the TTO valuation of UK EQ-5D (Dolan et al., 1996), so that TTO values for states worse than dead were bounded by -1. The interviewer booklet, which shows the details of the protocol used in the valuation survey of CORE-6D health states, is provided in Appendix 9.

Use of the TTO method and the MVH protocol in particular was dictated by NICE guidance on the methodology that should be adopted for the evaluation of technologies in its appraisal programme. According to this guidance, "*when EQ-5D data are not available or are inappropriate for the condition or effects of treatment, the valuation methods should be fully described and comparable to those used for the EQ-5D*". More specifically, "*the valuation of descriptions should use the time trade-off method in a representative sample of the UK population, with 'full health' as the upper anchor, to retain methodological consistency with the methods used to value the EQ-5D*" (National Institute for Health and Clinical Excellence, 2008). The rationale behind this guidance was to ensure comparability across the Institute's Appraisal programme. Additional guidance from the NICE Decision Support Unit specified that "*comparability with EQ-5D is enhanced by using the same valuation technique* [i.e. TTO], *the same variant of the technique* [i.e. use of visual props] *and by the same mode* [i.e. interviewer-administered] (Brazier & Rowen, 2011). Therefore, the protocol used for the valuation of CORE-6D was fully consistent with NICE guidance.

Interviews were conducted by trained and experienced interviewers from the Centre for Health and Social Care Research at Sheffield Hallam University. Valuations were elicited from members of the UK general public, as recommended by NICE (National Institute for Health and Clinical Excellence, 2008). Respondents were selected using sampling from streets in both urban and rural areas with a mix of socio-demographic characteristics in the North of England using a comprehensive contact management system for names and addresses in the UK (AFD Names and Numbers version 3.1.25 database, AFD Software Limited, Ramsey, UK). Households in these areas received letters informing them that interviewers would be in their area and interviewers then visited houses. Subsequently, all eligible and willing participants were interviewed in the respondent's own home. Eligible population consisted of adults aged over 18 years, who were considered by the interviewers to be cognitively able to participate in an interview. Addresses were visited up to four times on different days and times of the day before an address was considered a non-responder. No financial reward was offered for participation in the survey.

Ethical approval for the survey was received by the School of Health and Related Research (ScHARR) Research Ethics Committee at the University of Sheffield, as part of a wider MRC-NIHR funded methodology project (Condition-Specific Methodology for estimating QALYs: Developing and testing methods for deriving preference-based measures of health from condition specific measures - CoSMeQ) (Brazier et al., 2012). The letter from the ScHARR Research Ethics Committee confirming ethical approval for this project is provided in Appendix 10.

Funding for the valuation survey was provided by the MRC-NIHR Methodology Research Programme (project number 06/97/04). The funding received for the valuation survey was sufficient for 225 interviews. Previous valuation exercises have shown that respondents cannot value more than 13 health states during an interview (Dolan et al., 1996), and typically they are asked to value between 6 and 8 health states (Brazier et al., 2002, 2005b & 2008; Dolan et al., 1996; Yang et al., 2011). In order to increase the number of health states valued in a

survey, respondents can be divided into smaller sub-groups, each valuing different health states. One health state needs to be valued by all respondents, to allow comparison of mean values elicited from different sub-groups. For the valuation of CORE-6D, the 225 respondents were divided into 3 sub-groups, each provided with a card block of 8 health state cards; every card described a different health state, with the exception of the state that was valued by all respondents, which was shown on 3 cards distributed across the 3 card blocks, respectively. This arrangement allowed valuation of 22 health states, including the health state valued by all sub-groups.

Respondents were first asked to self-complete EQ-5D and CORE-6D for their own health, so as to become familiarised with the idea of describing health states, as well as with the items and response levels of CORE-6D. Subsequently, each respondent was given one of the three card blocks and undertook warm-up ranking and TTO tasks. This task, which allowed respondents to become familiarised with the cards and with the notion of having preferences for one health state over another, was followed by TTO valuations of 8 CORE-6D health states. If, during the TTO valuations, it was made clear that a respondent did not understand the TTO task, the interview was terminated by the interviewer and these partially completed interviews were not included in the dataset for analysis. The following exclusion criteria were applied: respondents with two or fewer responses; respondents who valued the worst state higher than all other states; respondents who valued all states worse than being dead; and respondents who valued all states identically but lower than 1.

Each interviewer started with a different card block with their first respondent, and moved on systematically alternating card blocks in the same order in successive interviews, e.g. the interviewer starting with card block 1 for the first respondent moved to card block 2 with the second respondent, then used card block 3 for the third respondent, then back to 1 with the fourth respondent, and so on. Because of the nature of some item responses (e.g. I make plans to end my life), respondents were informed in the cover letter and information sheet that the interview was about common mental and physical health

problems. In the information sheet and in a ‘thank you note’ left at the end of the interview all respondents were strongly recommended that they seek appropriate professional support either from their general practitioner (GP) or from a professional agency such as the Samaritans (their contact details being provided) if the interview raised personal issues for them. Finally, respondents were asked a number of background questions covering health, demographic and socio-demographic characteristics and how difficult they found the valuation tasks. The self-completion booklet provided to participants in the survey is presented in Appendix 11.

6.2.3 Modelling health state values using Rasch analysis

The standard approach for modelling utility values for health states described by a health state classification has been by creating dummy variables for each level of every dimension of an instrument (Brazier et al., 2002; Dolan, 1997) and regressing these onto the health state values (obtained using TTO or SG). However, this approach was not appropriate in the case of CORE-6D, since the highly correlated items of its emotional component were expected to produce significant, multiple interaction effects, and consideration of all possible interactions across different response levels of different items would require complex regression models as well as valuation of a large number of health states in order to predict utility values for all health states of the instrument (Brazier et al., 2007). An alternative method for modelling utility values derived from unidimensional PBMs has been described by Young and colleagues (2010): this method, which is based on Rasch analysis, employs a series of regression analyses in order to explore the relationship between the utility values derived from a valuation survey and the respective Rasch model logit values of the health states included in the survey. The selected regression model that best defines this relationship is then used to predict utility values for all potential states of the unidimensional PBM.

Nevertheless, this new method alone was not adequate for the estimation of utility values for CORE-6D; this is because CORE-6D is a 2-dimensional scale, consisting of a unidimensional emotional component and a physical item. Thus, in order to predict utility values for all health states described by CORE-6D taking into account the effect of the physical item, a hybrid approach was

adopted: the methodology described by Young and colleagues (2010), which can be used for the prediction of utility values in the case of unidimensional measures such as the emotional component of CORE-6D, was combined with the standard approach used for the prediction of utility values that relies on the use of dummy variables to reflect all the different levels of each dimension; in the case of CORE-6D, the dummy variables represented the different severity levels of the physical item.

More specifically, a series of regression analyses were undertaken to explore the relationship between the utility value of each health state of CORE-6D that was considered in the valuation survey and

a. the respective Rasch model logit value corresponding to the emotional component of the health state, as calculated by previously undertaken Rasch analysis of CORE-6D data on [N400a]
b. the response level (0, 1 or 2) of the physical item of the health state, modelled in the form of 2 dummy dichotomous variables, one for response level 1 and one for response level 2.

OLS models were used to analyse the valuation data at an aggregate (mean) level first, i.e. regression analyses were carried out on the mean utility values obtained for each of the 18 health states included in the valuation survey, without taking into account individual respondent characteristics (such as age, gender, ethnicity, etc.), since aggregate utility data (i.e. data at the population level) are typically those that are used in cost-utility analyses of healthcare technologies. Previous research has shown that, despite having fewer degrees of freedom available for analysis, aggregate models may perform equally or even better than individual-level ones in predicting mean health state utility values (Brazier et al., 2002 & 2008; McKenna et al., 2008; Yang et al., 2009 & 2011).

Various regression models were fitted on the data, including simple linear, quadratic and cubic forms, to reflect potential non-linearities in the relationship between the utility values (dependent variable) and the Rasch model logit

scale. These ‘base-case’ models assumed an additive relationship between the emotional component of CORE-6D and the physical item of the measure. Moreover, models that took into account the potential (multiplicative) interaction between the emotional component of CORE-6D and the physical item (also considering linear, quadratic and cubic relationship) were tested to explore whether considering multiplicative interactions between the two dimensions of CORE-6D improved the overall model fit; for this purpose, interaction variables were added to the best solution identified among the base-case models. The model fit and predictive ability was assessed using the coefficient of determination (adjusted R-Squared), the root mean squared error (RMSE), the mean absolute error (MAE) [i.e. the mean absolute difference between the predicted and the observed utility value across all health states], and the number of health states with absolute error above 0.01, 0.05 and 0.10 (Brazier et al., 2007). The model with the best fit at the mean level was selected in order to predict mean TTO values for all health states described by CORE-6D based on their respective Rasch model logit value and the response level of the physical item. In order to test the fit of the selected model, residuals of all health states (i.e. differences between predicted and observed utility values) were plotted against each of the independent variables of the model, as well as against the predicted utility values, to confirm lack of any systematic relationship (Altman, 1991). The predictive ability of the selected model was also assessed by visually inspecting the plot of the predicted utility values against the observed utility values that were obtained in the valuation survey.

In addition, OLS regression analyses at the individual level were carried out, to explore the impact of respondents’ personal characteristics including age, gender, ethnicity, marital status, home ownership, level of education and employment status, on the elicited utility values. An important limitation of the OLS model is that it assumes a continuous variable without censoring; in this case, it does not allow for the dependent variable (utility value) to be bounded by a maximum value of +1 and a minimum value of -1. Therefore, Tobit models were estimated, which allowed censoring at both the top and bottom ends of the relationship (Tobin, 1958). The general Tobit model with upper and lower censoring limit of +1 and -1, respectively, is defined as:

$$y^*_{ij} = \beta x_i + \theta r_i + \delta z_{ij} + \varepsilon_{ij}$$

where y^*_i the unobservable latent variable. The censored observed outcome y_i is:

$$y_i = \begin{cases} -1 & \text{if } y^*_i \leq -1 \\ y^*_i & \text{if } -1 < y^*_i < 1 \\ 1 & \text{if } y^*_i \geq 1 \end{cases} \qquad \text{(Long, 1997)}$$

The fit of individual OLS models was assessed using the adjusted R-Squared and the RMSE. Tobit models were assessed using the estimated standard error of the regression, which is analogous to the RMSE in OLS regression.

OLS mean-level analyses were performed on SPSS version 19 (IBM Corp., 2010); all individual-level analyses (OLS and Tobit models) were run on STATA version 10 (Stata Corp., 2007).

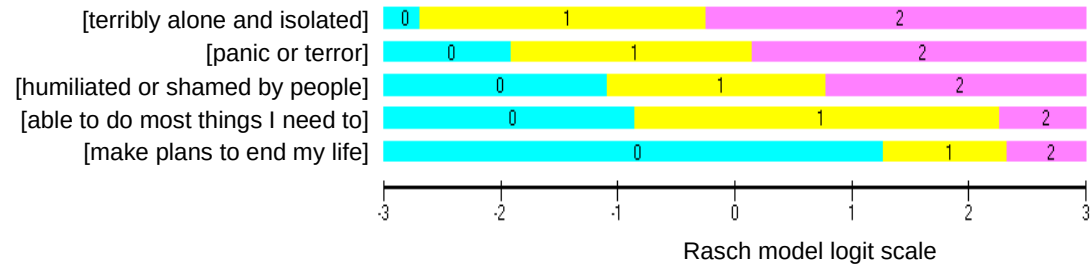
6.3 Results

6.3.1 Generation of plausible health states using Rasch analysis

Identification of plausible health states from the emotional component of CORE-6D

Identification of plausible health states from the emotional component of CORE-6D was achieved by inspection of the Rasch item threshold map that was produced at the development of the emotional component of CORE-6D, as reported in Chapter 5 (section 5.5). The item threshold map of the emotional component of CORE-6D is illustrated in Figure 16.

Figure 16. Rasch item threshold map of the emotional component of CORE-6D



0 = never; 1 = only occasionally or sometimes; 2 = often, most or all the time; note that the fourth item is positively worded and therefore response levels are reversed

The map depicts the most likely combinations of responses to the 5 items of the emotional component of CORE-6D across the continuum of the emotional symptom severity. Items of the emotional component have been ordered from the easiest to the most difficult one as indicated by their mean location on the Rasch logit scale. Coloured areas 0 (light blue), 1 (yellow) and 2 (purple) correspond to the 3 response levels of the measure, that is, ‘never’, ‘only occasionally or sometimes’, and ‘often, most or all the time’, respectively, with the exception of the positively worded item (I am able to do most things I need to), the response levels of which are reversed. The map allows prediction of the most likely responses at each level of emotional symptom severity captured by the interval Rasch scale. For example, a person whose level of emotional distress corresponds to a Rasch logit value of +1 is expected to most likely respond 22210 to the 5 items of the emotional component of CORE-6D, ordered from the easiest to the most difficult one, 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 28, 11 distinct emotional health states were identified along the Rasch model logit scale, each reflecting the most likely emotional state to be observed in a person with common mental health problems at a specific level of emotional symptom severity. Detailed descriptions of these 11 health states depicted in the Rasch item threshold map are provided in Appendix 12.

Table 28. Plausible health states of the emotional component of CORE-6D as identified by the Rasch item threshold map and frequency of each health state in the study sample [N400a]

Item	Health states										
	1	2	3	4	5	6	7	8	9	10	11
[terribly alone and isolated]	N	S	S	S	S	O	O	O	O	O	O
[panic or terror]	N	N	S	S	S	S	O	O	O	O	O
[humiliated or shamed by people]	N	N	N	S	S	S	S	O	O	O	O
[able to do most things I need to]	O	O	O	O	S	S	S	S	S	N	N
[make plans to end my life]	N	N	N	N	N	N	N	N	S	S	O

N = never; S = only occasionally or sometimes; O = often, most or all the time; note that the 4th item is positively worded and therefore response levels are reversed

The emotional component of CORE-6D can describe $3^5 = 243$ emotional health states. The 11 emotional states identified by inspection of the Rasch item threshold map represent only 4.5% of the health states that can be described by the emotional component. However, these 11 states covered 37.1% of the response combinations obtained from the study sample [N400a], that is, from the random study sample used in Rasch analysis of CORE-OM data, and 33.7% of the response combinations observed in [N1500], which is the original dataset the [N400a] was derived from, after excluding cases with one or more responses missing. The frequency and percentage of individuals in [N400a] and [N1500] that experienced each of the 11 emotional health states among those that provided complete responses to the emotional component of CORE-6D are shown in Table 29.

In contrast, as it can be seen in Table 30, the coverage of the 15 health states derived using an orthogonal block design (generated on IBM SPSS Statistics 19) on the full range of emotional health states described by CORE-6D was only 14.5% in [N400a] and 14.1% in [N1500]. Moreover, some of the states generated using the latter approach were not plausible, 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’.

Table 29. Frequency and percentage of observations of the 11 emotional health states of CORE-6D that were identified by inspection of the Rasch item threshold map in datasets [N400a] and [N1500]

Emotional state	*Classification*	[N400a]		[N1500]	
		Frequency	%	Frequency	%
State 1	**00000**	18	5.3	73	5.6
State 2	**10000**	20	5.9	82	6.3
State 3	**11000**	21	6.2	60	4.6
State 4	**11100**	17	5.0	54	4.1
State 5	**11110**	19	5.6	69	5.3
State 6	**21110**	9	2.7	29	2.2
State 7	**22110**	9	2.7	29	2.2
State 8	**22210**	5	1.5	24	1.8
State 9	**22211**	5	1.5	14	1.1
State 10	**22221**	0	0.0	3	0.2
State 11	**22222**	2	0.6	5	0.4
Total in the 11 health states		125	37.1	442	33.7
Total number of complete observations		337		1310	

Each emotional state is represented by a five digit code ('classification') that indicates the response level of each of the 5 emotional items, from left to right: I feel terribly alone and isolated; I feel panic or terror; I feel humiliated or shamed by other people; I am able to do most things I need to; I make plans to end my life.

Table 30. Frequency and percentage of observations of the 15 emotional health states of CORE-6D that were identified using an orthogonal design, in datasets [N400a] and [N1500]

Emotional state	*Classification*	[N400a]		[N1500]	
		Frequency	**%**	**Frequency**	**%**
State 1	00000	18	5.3	73	5.6
State 2	00120	0	0.0	0	0.0
State 3	00210	0	0.0	0	0.0
State 4	01022	0	0.0	0	0.0
State 5	01201	0	0.0	0	0.0
State 6	02012	0	0.0	0	0.0
State 7	02101	0	0.0	0	0.0
State 8	10021	0	0.0	0	0.0
State 9	10202	0	0.0	0	0.0
State 10	11110	19	5.6	69	5.3
State 11	12000	2	0.6	12	0.9
State 12	20011	0	0.0	0	0.0
State 13	20102	0	0.0	0	0.0
State 14	21000	5	1.5	19	1.5
State 15	22220	5	1.5	12	0.9
Total in the 15 health states		49	14.5	185	14.1
Total number of complete observations		337		1310	

Each emotional state is represented by a five digit code ('classification') that indicates the response level of each of the 5 emotional items, from left to right: I feel terribly alone and isolated; I feel panic or terror; I feel humiliated or shamed by other people; I am able to do most things I need to; I make plans to end my life.

In order to obtain full CORE-6D health states, each emotional health state needs to be combined with different response levels of the physical item. The 11 emotional health states selected by inspection of the Rasch item threshold map combined with the 3 response levels of the physical item of CORE-6D produce a 2-dimensional set of 11 x 3 = 33 health states that are overall frequently observed in the study population and, as such, are plausible.

Selection of plausible health states for the valuation survey

As reported in section 6.2.2, the number of respondents and their arrangement in 3 sub-groups allowed the valuation of 22 health states (with 1 health state being valued by all 3 sub-groups). Selection of the 22 health states of CORE-6D for consideration in the valuation survey was made in collaboration with Dr

Donna Rowen, research fellow at ScHARR, University of Sheffield, who has expertise in this field. As described earlier, 33 full CORE-6D health states were constructed by combining the 11 emotional health states identified by inspection of the Rasch item threshold map with the 3 response levels of the physical item. However, the emotional health state 10 (22221) was not represented in the study sample [N400a] (as shown in Table 29), had a very narrow logit range (as shown on the item threshold map in Figure 16) and was therefore excluded from further consideration, leaving 30 full CORE-6D health states as candidates for inclusion in the valuation survey. In addition, as part of the CoSMeQ study (Brazier et al., 2012) a number of emotional health states without any reference to the physical component were also selected for valuation, so as to assess the impact of the addition of the physical component on valuations of the emotional states of CORE-6D, and this allowed fewer full health states to be valued (since the limit was the valuation of 22 health states in total).

Selection of the 22 health states for the valuation survey was made as follows: First, the 10 emotional health states chosen using the Rasch vignette approach were combined with the physical item at response level zero (never troubled by aches, pains, or other physical problems) and were included in the valuation survey. In addition, and in order to assess the impact of physical functioning on utility values, 4 of these emotional states (including best state 00000, worst state 22222 and two intermediate states) were combined with levels 1 and 2 of the physical item, so as to cover the full severity range captured by CORE-6D, thus producing another 8 CORE-6D health states. The criteria for selecting the two intermediate emotional states for valuation were as follows:

- relative frequency of the state in the study samples [N400a] and [N1500] (as shown in Table 29) – states with high frequency were preferred
- location coverage (range) of the state on the item threshold map (shown in Figure 16) – states with wider location coverage were favoured

- relative distance between the 4 states (best, worse and the 2 intermediate) – ideally the selected states should be of variant symptom severity

Based on the above criteria, intermediate emotional states 3 (11000) and 7 (22110) were selected for combination with response levels 1 and 2 of the physical item and inclusion in the valuation survey. In addition to the 18 full health states, 4 emotional health states (best 00000, worst 22222 and intermediate states 11000 and 22110 as chosen already) with no reference to the physical item were also selected for use in the related CoSMeQ study (Brazier et al., 2012). Responses to the states describing only the emotional component of CORE-6D were analysed separately and are available in the study by Brazier and colleagues (2012). A sample of a health state card used in the valuation survey is presented in Table 31.

Table 31. Sample of a health state card used in the valuation survey – card describing CORE-6D state 221101

- You feel terribly alone and isolated often, most or all the time
- You feel panic or terror often, most or all the time
- You feel humiliated or shamed by other people only occasionally or sometimes
- You are able to do most things you need to only occasionally or sometimes
- You never make plans to end your life
- You are troubled by aches, pains or other physical problems only occasionally or sometimes

The 22 health states were distributed across the 3 card blocks used in the valuation survey so that each person was asked to value a variety of health states across the range of symptom severity captured by CORE-6D. The health states contained in each of the card blocks are presented in Table 32. Two of the card blocks contained 8 full CORE-6D health states each. The

other card block contained 4 full CORE-6D health states, and 4 emotional health states (identical with the emotional components of the 4 full CORE-6D states already included in this card block, but without any reference to the physical item). CORE-6D state 222220 was included in all 3 card blocks. All respondents first ranked and valued 4 states and subsequently ranked and valued the remaining 4 states in the card block. In the card block that contained 4 emotional health states without reference to the physical item and 4 full CORE-6D states, the emotional states were ranked and valued first, followed by ranking and valuation of the full CORE-6D states, so that responders were not aware of the presence of the physical item when valuing the 4 emotional states; in the other two card blocks, the 4 full CORE-6D states that were ranked and valued first were chosen at random.

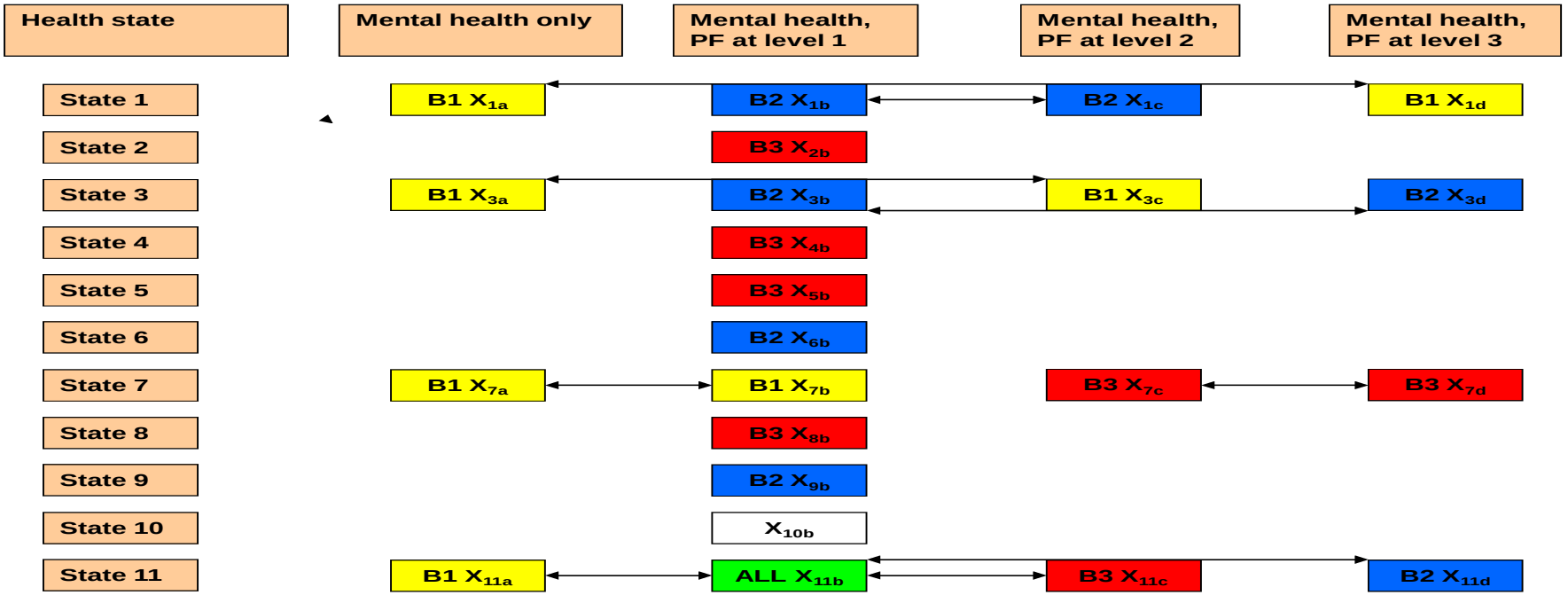
Table 32. Health states included in each of the 3 cardblocks used in the valuation survey of CORE-6D

Card block 1	Card block 2	Card block 3
00000N	000000	100000
11000N	000001	111000
22110N	110000	111100
22222N	110002	221101
000002	211100	221102
110001	222110	222100
221100	222222	222221
222220	222220	222220

Each health state is represented by a six digit code that indicates the response level of each of the 5 emotional items plus the physical item, from left to right: I feel terribly alone and isolated; I feel panic or terror; I feel humiliated or shamed by other people; I am able to do most things I need to; I make plans to end my life; I feel aches, pains, or other physical problems. Response levels for the positively worded item are reversed. 'N' signifies that the physical item was not included in these states and therefore was not mentioned in the card.

Similarly, Figure 17 shows the allocation of states in each card bloc: states in card block 1 [B1] are coloured yellow; states in card block 2 [B2] are coloured blue; states in card block 3 [B3] are coloured red; the state coloured green was included in all 3 card blocks. The figure helps identify states that share the same level of emotional component but differ in the response level of the physical item and vice versa, thus allowing exploration of the impact of the level of physical functioning on the health state utility value.

Figure 17. Health states included in each of the three card blocks used in the valuation survey of CORE-6D



PF = physical functioning; card bloc 1 includes the states highlighted in yellow, card bloc 2 includes the states highlighted in blue and card bloc 3 includes the states highlighted in red. State highlighted in green was valued by all respondents in the survey. Each health state box includes the bloc (B1, B2, B3 or ALL), the mental health state number and the level of severity of the physical functioning dimension. For example, x_{1a} indicates mental health state 1 and no mention of physical functioning, x_{3b} indicates mental health state 3 with physical functioning at level 0, x_{5c} indicates mental health state 5 with physical functioning at level 1 and x_{7d} indicates mental health state 7 with physical functioning at level 2. The arrows indicate how the states can be used to estimate the relationship between utility and physical functioning severity.

6.3.2 Valuation survey

Respondents' characteristics

The valuation survey, which was conducted on 225 respondents, achieved a response rate of 45.7% for respondents answering their door at the time of interview. The study achieved a completion rate of 99.7% for all 18 health states included in the valuations of full CORE-6D states that were considered in this thesis (only 4 utility values were missing). Characteristics of all respondents included in the analysis are presented in Table 33, which allows comparison of the valuation study sample to the general population in South Yorkshire and England (Kind et al., 1999). The study sample had a higher mean age, a higher proportion of females, home owners and retired individuals, and a lower proportion of employed/self-employed individuals.

A large proportion of respondents reported that they found the rank (35.1% of respondents) and TTO (40.9% of respondents) tasks either 'very difficult' or 'rather difficult', and this likely includes both respondents who found completion of the task complex and respondents who found the decisions involved challenging. Finding a task difficult does not convey a lack of understanding, as no respondents met the set exclusion criteria (i.e. providing ≤ 2 responses; valuing the worst state higher than all other states; valuing all states worse than being dead; or valuing all states identically but lower than 1) that indicated no understanding of the TTO task. Moreover, interviewers reported that it was doubtful (according to their expert judgment) whether the respondent understood the rank and TTO tasks in just 5.8% and 4.9% of the interviews, respectively.

Table 33. Characteristics of respondents in the valuation survey and comparison with population characteristics for South Yorkshire and England

Variable	Respondents (n=225)	South Yorkshire[1]	England[1]
Mean age (SD)	48.9 (17.2)	-	-
Age distribution			
18-40	32.7%	41.2%	41.6%
41-65	48.0%	39.1%	39.1%
Over 65	19.3%	19.7%	19.3%
Female	58.7%	51.2%	51.3%
Married/Partner	69.8%	-	-
Employed or self-employed	51.3%	56.1%	60.9%
Unemployed	3.1%	4.1%	3.4%
Long-term sick	5.4%	7.7%	5.3%
Full-time student	5.4%	7.5%	7.3%
Retired	22.3%	14.4%	13.5%
Own home outright or with a mortgage	80.0%	64.0%	68.7%
Renting property	20.0%	36.0%	31.3%
Secondary school is highest level of education	37.9%	-	-
Average EQ-5D score (SD)	0.83 (0.28)	-	0.86 (0.23)[2]
TTO completion rate	99.7%	-	-
Respondent found 1st rank valuation task very or rather difficult	35.1%	-	-
Respondent found 1st TTO valuation task very or rather difficult	40.9%	-	-
Interviewer doubted whether respondent understood 1st rank task	5.8%	-	-
Interviewer doubted whether respondent understood 1st TTO task	4.9%	-	-

1. Statistics for South Yorkshire Health Authority and for England in the Census 2001. Questions used in this study and the census are not identical. The census includes persons aged 16 and above whereas this study surveyed persons aged 18 and above only. Age distribution is here reported as the percentage of all adults aged 18 and over.
2. Interviews conducted in the Measurement and Valuation of Health (MVH) study (Kind et al., 1999).
SD = standard deviation

Health state utility values

The descriptive statistics for the utility values obtained for each health state valued in the survey are reported in Table 34. The mean utility values range from 0.96 (best state 000000) to 0.10 (worst state 222222) and have large standard deviations. Median values are consistently higher than mean ones (with the exception of worst state 222222 where the median and mean utility value are equal), indicating a negative skewness of the data. This skewness is most apparent in the histogram created from the utility values obtained in the survey, which is illustrated in Figure 18. The histogram reveals that a substantial proportion of responses (466/1492, i.e. 31.2%) corresponded to a utility value of 1.0, illustrating that on many occasions respondents were not prepared to sacrifice time for quality of life. Overall, the results of the survey indicate 3 types of respondents: a. a small proportion of respondents (14/225, i.e. 6.2%) who never trade time for quality (and thus attach a utility value of 1.0 to any state); b. a significant proportion of respondents (70/225, i.e. 31.1%) who always trade time for quality (and thus never attach a utility value of 1.0 to any state) and c. the largest proportion of respondents (141/225, i.e. 62.7%) who do both, depending on the state valued (so that they attach a utility value of 1.0 to some states and a utility value less than 1.0 to other states).

Table 34. Utility values by CORE-6D health state obtained in the valuation survey

CORE-6D health state	Utility value								
	N	Mean	SD	Minimum	25th percentile	Median	75th percentile	Maximum	Mode
000000	74	0.96	0.13	0.08	0.99	1.00	1.00	1.00	1.00
000001	75	0.93	0.14	0.33	0.93	1.00	1.00	1.00	1.00
000002	76	0.82	0.32	-0.93	0.78	0.93	1.00	1.00	1.00
100000	74	0.87	0.22	0.08	0.84	1.00	1.00	1.00	1.00
110000	75	0.88	0.25	-0.73	0.85	1.00	1.00	1.00	1.00
110001	76	0.86	0.27	-0.93	0.80	0.96	1.00	1.00	1.00
110002	75	0.74	0.31	-0.83	0.57	0.83	1.00	1.00	1.00
111000	74	0.79	0.29	-0.23	0.69	0.93	1.00	1.00	1.00
111100	74	0.76	0.33	-0.40	0.53	0.93	1.00	1.00	1.00
211100	75	0.66	0.35	-0.63	0.50	0.73	1.00	1.00	1.00
221100	75	0.57	0.44	-0.93	0.45	0.63	0.93	1.00	1.00
221101	73	0.49	0.47	-0.88	0.30	0.50	0.88	1.00	1.00
221102	74	0.40	0.49	-0.93	0.14	0.44	0.83	1.00	1.00
222100	74	0.47	0.43	-0.93	0.20	0.50	0.84	1.00	1.00
222110	74	0.38	0.45	-0.98	0.08	0.44	0.70	1.00	1.00
222220	225	0.23	0.52	-0.98	0.00	0.30	0.53	1.00	1.00
222221	74	0.21	0.50	-0.93	-0.08	0.23	0.50	1.00	1.00
222222	75	0.10	0.53	-0.93	-0.33	0.10	0.48	1.00	1.00

SD = standard deviation

Figure 18. Histogram of the utility values obtained in the valuation survey of CORE-6D health states

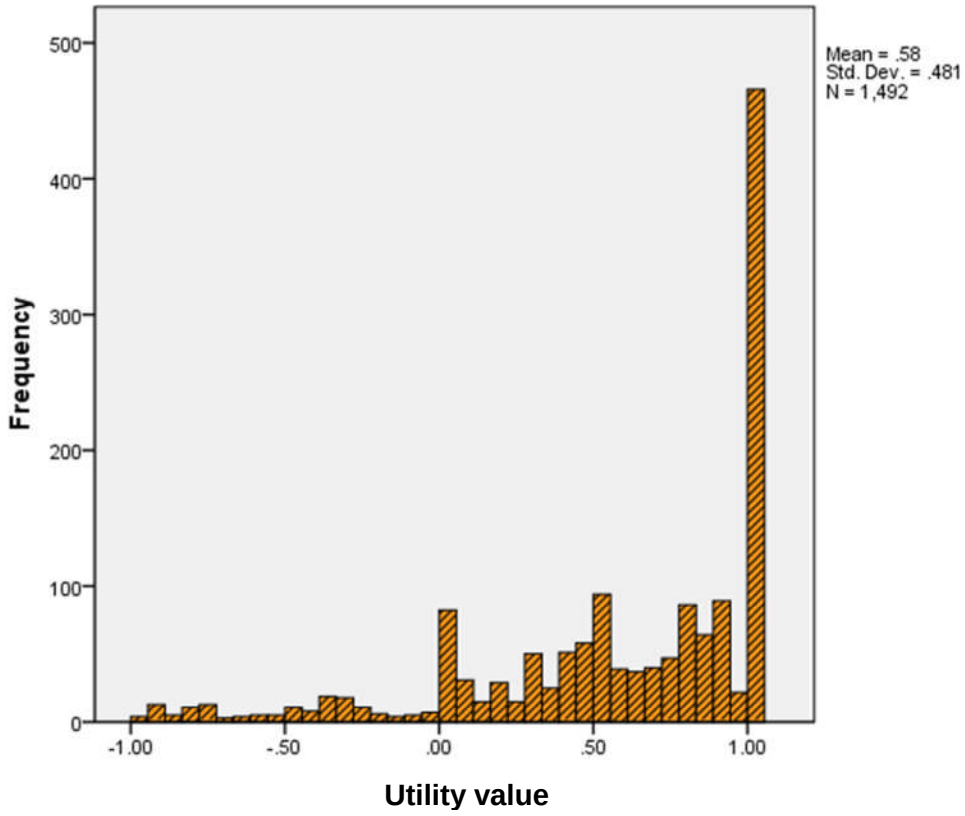


Table 35, which shows responses by card block, demonstrates the changes in obtained mean utility values with increasing severity of physical and emotional symptoms: moving to states with more severe physical symptoms (i.e. increasing the response level of the physical item), while keeping the emotional health state unchanged, results in a decrease in the mean utility value; similarly, moving to states with more severe emotional symptoms (i.e. moving from emotional state 00000 to emotional state 22222), while keeping the response level of the physical item intact, also results in a decrease in the mean utility value. There is only one inconsistency to this pattern, observed for states 100000 and 110000; in this case the mean utility value increased by a small and non-significant amount (from 0.87 to 0.88, respectively) despite of the increase in the emotional symptom severity. This inconsistency can be explained by the fact that these health states were included in different card blocs and hence were valued by different respondents.

Table 35. Mean utility values for each CORE-6D health state included in valuation survey by severity of emotional and physical symptoms

CORE-6D	Response levels of the physical item		
Emotional component	0	1	2
00000	0.96 (0.13)	0.93 (0.14)	0.82 (0.32)
10000	0.87 (0.22)		
11000	0.88 (0.25)	0.86 (0.27)	0.74 (0.31)
11100	0.79 (0.29)		
11110	0.76 (0.33)		
21110	0.66 (0.35)		
22110	0.57 (0.44)	0.49 (0.47)	0.40 (0.49)
22210	0.47 (0.43)		
22211	0.37 (0.45)		
22221			
22222	0.23 (0.52)	0.21 (0.50)	0.10 (0.53)

Health states included in each card bloc are highlighted in a different colour: states in card bloc 1 are highlighted in yellow, states in card bloc 2 are highlighted in blue and states in card bloc 3 are highlighted in red; all respondents valued state 222220, highlighted in green; standard deviation is provided in parenthesis.

6.3.3 Modelling health state values using Rasch analysis

Mean-level models

A number of OLS models at the mean level were explored using as independent (explanatory) variables the Rasch model logit value (assuming simple linear, quadratic and cubic relationships) as suggested by Young and colleagues (2010) and 2 dummy variables accounting for the response level of the physical item. The models aimed to predict utility values for the 33 CORE-6D health states that are formed by combining the emotional states depicted in the Rasch item threshold map with the 3 response levels of the physical item. However, given that emotional health states with the same total (ordinal) score correspond to the same Rasch logit value, it is possible to predict utility values for all CORE-6D health states, based on their total emotional component score and the response level of the physical item.

The Rasch model logit values for each emotional state identified on the Rasch item threshold map were rescaled and anchored at 0.96 and 0.23, which were the observed mean utility values obtained in the valuation survey for the best state 00000 and worst state 22222, respectively. Rescaling was achieved using the formula:

$$z_i = max_{new} + r * (min_{Rasch} * x_i)$$

where z_i is the Rasch model rescaled logit value of emotional state *i*, x_i is the Rasch model original logit value of the emotional state *i*, max_{new} is the maximum value of the new scale, min_{Rasch} is the minimum value of the Rasch original logit scale, and r is the range of the new scale divided by the range of the Rasch original logit scale. This process did not alter the interval scale properties of the Rasch model logit values, but converted the original Rasch scale (which ranged from -3.748 to +3.562) into a more easily interpretable scale (ranging from 0.23 to 0.96).

The following model specifications were tested:

Model m1 – simple linear relationship: $y = \alpha + \beta_1 R + \gamma_1 P_1 + \gamma_2 P_2$

Model m2 – quadratic relationship: $y = \alpha + \beta_2 R^2 + \gamma_1 P_1 + \gamma_2 P_2$

Model m3 – cubic relationship: $y = \alpha + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$

Model m4 – quadratic relationship: $y = \alpha + \beta_1 R + \beta_2 R^2 + \gamma_1 P_1 + \gamma_2 P_2$

Model m5 – cubic relationship: $y = \alpha + \beta_1 R + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$

Model m6 – cubic relationship: $y = \alpha + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$

Model m7 – cubic relationship: $y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$

where y is the mean predicted utility value, R is the Rasch model rescaled logit value, P_1 is a dummy variable for response level 1 of the physical item (I have been troubled by aches, pains, physical problems only occasionally or sometimes), P_2 is a dummy variable for response level 2 of the physical item (I have been troubled by aches, pains, physical problems often, most or all the time), α is the constant, and β_i and γ_i are regression coefficients.

The regression coefficients and goodness of fit statistics for all 7 base-case models are shown in Table 36. The adjusted R-squared statistics varied from 0.773 (model m3) to 0.990 (model m7); these high values may be an artefact of the the relatively large number of independent variables (ranging from 3 to 5) compared with the number of mean obsservations (n=18). It needs to be noted that the adjusted R-squared increased with increase of independent variables (from 3 to 5), indicating that the addition of extra variables improved the regression model more than what would be expected by chance. The RMSE ranged from 0.0275 to 0.1292, while the MAE ranged from 0.014 (standard deviation 0.019) to 0.102 (standard deviation 0.052). In two models (m2 and m3) a number of states demonstrated absolute errors higher than 0.10 (5 and 9 states, respectively); in no model were absolute errors for all states lower than 0.05, but in model m7 the absolute error of 17 out of the 18 health states was lower than 0.05 (one state had an absolute error between 0.05 and 0.10). The regression coefficients of all independent variables were overall statistically significant at the 0.05 level across the models; only dummy variable P_1 was non-significant in any of the models, although in model m7 the level of significance was only moderately above 0.05 (0.07).

The largest mean-level model with linear, quadratic and cubic terms for the Rasch logit value and 2 physical dummy variables (model m7) had the highest R-squared (0.990), the lowest RMSE (0.0275), the lowest MAE (0.014) and the best predictive ability (14 out of the 18 health states showed absolute error no more than 0.01, 3 health states had absolute error above 0.01 and up to 0.05, and only one state had an absolute error of 0.07). The constant and the regression coefficients of all independent variables were statistically significant, with the exception of P_1, the level of significance of which reached 0.07. Based on these findings, model m7 was the preferred solution among the base-case models for the prediction of utility values of the 33 CORE-6D health states.

Additional models that considered multiplicative interaction terms[3] between the emotional component and the physical item of CORE-6D did not appear to offer any improvement in the model fit compared with the selected model m7. As it can be seen in Table 37, in none of these additional models were the interaction terms significant. Moreover, consideration of interaction terms did not offer any improvement in the model fit and its predictive ability, as suggested by the RMSE and MAE statistics and the adjusted R-Squared values, which were, at best, equivalent to those of the selected mean-level additive OLS model m7. These findings suggest that a simple additive model was adequate to capture the relationship between the utility values on the one side of the equation, and the Rasch model logit value of the emotional component as well as the physical dummy variables on the other.

[3] In order to estimate multiplicative interaction terms between the emotional component and the physical item of each health state, the response levels of each of the 6 items were rescored on a scale from 1-3 (instead of the original 0-2 scale). The total new score of the emotional component, as obtained by summing the individual emotional item scores, was multiplied by the new score of the physical item at the state level to give the multiplicative interaction term.

Table 36. Results of base-case mean-level ordinary least squares regression models for the prediction of CORE-6D utility values [analysis on N = 18 mean utility values]

Model	*Cons* (*p val*)	*R* (*p val*)	R^2 (*p val*)	R^3 (*p val*)	P_1 (*p val*)	P_2 (*p val*)	Adj R-sq	RMSE	MAE (SD)	No. >|0.01|	No. >|0.05|	No. >|0.10|
m1 $y = \alpha + \beta_1 R + \gamma_1 P_1 + \gamma_2 P_2$	0.008 (0.833)	1.057 (0.000)			-0.044 (0.189)	-0.151 (0.000)	0.961	0.0533	0.040 (0.029)	12	5	0
m2 $y = \alpha + \beta_2 R^2 + \gamma_1 P_1 + \gamma_2 P_2$	0.302 (0.000)		0.844 (0.000)		-0.070 (0.219)	-0.177 (0.006)	0.886	0.0916	0.073 (0.036)	17	13	5
m3 $y = \alpha + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$	0.416 (0.000)			0.779 (0.000)	-0.085 (0.284)	-0.193 (0.025)	0.773	0.1292	0.102 (0.052)	18	14	9
m4 $y = \alpha + \beta_1 R + \beta_2 R^2 + \gamma_1 P_1 + \gamma_2 P_2$	-0.130 (0.100)	1.585 (0.000)	-0.443 (0.056)		-0.029 (0.329)	-0.137 (0.000)	0.969	0.0478	0.037 (0.019)	16	3	0
m5 $y = \alpha + \beta_1 R + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$	-0.108 (0.072)	1.388 (0.000)		-0.282 (0.025)	-0.028 (0.329)	-0.135 (0.000)	0.972	0.0452	0.034 (0.020)	17	3	0
m6 $y = \alpha + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$	0.099 (0.002)		2.624 (0.000)	-1.758 (0.000)	-0.029 (0.170)	-0.137 (0.000)	0.985	0.0331	0.024 (0.016)	13	2	0
m7 $y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2$	0.366 (0.004)	-1.695 (0.022)	5.712 (0.000)	-3.446 (0.000)	-0.033 (0.069)	-0.141 (0.000)	0.990	0.0275	0.014 (0.019)	4	1	0

Notes: *Adj R-sq* = adjusted R-squared; *Cons* = constant; *R*: Rasch model rescaled logit values; P_1 and P_2: dummy variables accounting for response levels 1 and 2, respectively, of the physical item of CORE-6D; MAE = mean absolute error; RMSE = root mean squared error; sd = standard deviation; *p* val: *p* value

Table 37. Results of mean-level ordinary least squares regression models considering potential multiplicative interactions between the emotional component and the physical item of CORE-6D – additional independent variables added to the best-performing additive model among the base-case mean-level model specifications (Model m7)

[analysis on N = 18 mean utility values]

Base-case model m7 plus…	*Cons*	*R* (*p* val)	R^2 (*p* val)	R^3 (*p* val)	P_1 (*p* val)	P_2 (*p* val)	*I* (*p* val)	I^2 (*p* val)	I^3 (*p* val)	Adj R-Sq	RMSE	MAE (SD)	No. >|0.01|	No. >|0.05|	No. >|0.10|
…linear multiplicative interaction	0.415	-1.790	5.803	-3.493	-0.018	-0.111	-0.002			0.989	0.0281	0.018	9	1	0
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_1 I$	(0.009)	(0.023)	(0.001)	(0.000)	(0.531)	(0.044)	(0.528)					(0.015)			
…quadratic multiplicative interaction	0.386	-1.766	5.804	-3.491	-0.030	-0.131		-0.000		0.989	0.0285	0.017	5	1	0
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_2 I^2$	(0.008)	(0.027)	(0.001)	(0.000)	(0.155)	(0.001)		(0.701)				(0.017)			
…cubic multiplicative interaction	0.375	-1.730	5.760	-3.470	-0.033	-0.137			-0.000	0.989	0.0286	0.016	5	1	0
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_3 I^3$	(0.009)	(0.031)	(0.001)	(0.000)	(0.096)	(0.000)			(0.855)			(0.017)			
…quadratic multiplicative interaction	0.557	-1.582	5.250	-3.244	0.060	0.006	-0.015	0.000		0.990	0.0276	0.064	14	4	3
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_1 I + \delta_2 I^2$	(0.009)	(0.044)	(0.002)	(0.001)	(0.416)	(0.955)	(0.217)	(0.255)				(0.084)			
…cubic multiplicative interaction	0.493	-1.633	5.415	-3.321	0.035	-0.034	-0.009		0.000	0.990	0.0277	0.034	9	3	1
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_1 I + \delta_3 I^3$	(0.007)	(0.037)	(0.001)	(0.001)	(0.535)	(0.688)	(0.218)		(0.276)			(0.043)			
…cubic multiplicative interaction	0.411	-1.694	5.622	-3.416	0.002	-0.086		0.000	0.000	0.989	0.0280	0.041	15	5	0
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_2 I^2 + \delta_3 I^3$	(0.006)	(0.033)	(0.001)	(0.001)	(0.955)	(0.104)		(0.246)	(0.265)			(0.031)			
…cubic multiplicative interaction	0.493	-1.633	5.415	-3.321	0.035	-0.034	-0.009	I^2 *term*	0.000	0.990	0.0277	0.034	9	3	1
$y = \alpha + \beta_1 R + \beta_2 R^2 + \beta_3 R^3 + \gamma_1 P_1 + \gamma_2 P_2 + \delta_1 I + \delta_2 I^2 + \delta_3 I^3$	(0.007)	(0.037)	(0.001)	(0.001)	(0.535)	(0.688)	(0.218)	*REM*	(0.276)			(0.043)			

Notes: *Adj R-Sq* = adjusted R-squared; *Cons* = constant; *R*: Rasch model rescaled logit values; P_1 and P_2: dummy variables accounting for response levels 1 and 2, respectively, of the physical item of CORE-6D; *I* = multiplicative interaction term; δ_i = additional regression coefficients; REM = removed from regression due to multi-collinearity (i.e. strong correlation with other independent variables, which increases the standard errors of the coefficients); MAE = mean absolute error; RMSE = root mean squared error; sd = standard deviation; *p* val: *p* value

Figure 19 shows the plotting of residuals obtained from model m7 against each of the independent variables of the model, as well as against the predicted utility values. Despite the small number of data points, the plots suggest that the points are randomly scattered across x values indicating a good model fit.

Figure 19. Plots of residuals against (a) each of the five independent variables of selected model m7 and (b) predicted utility values

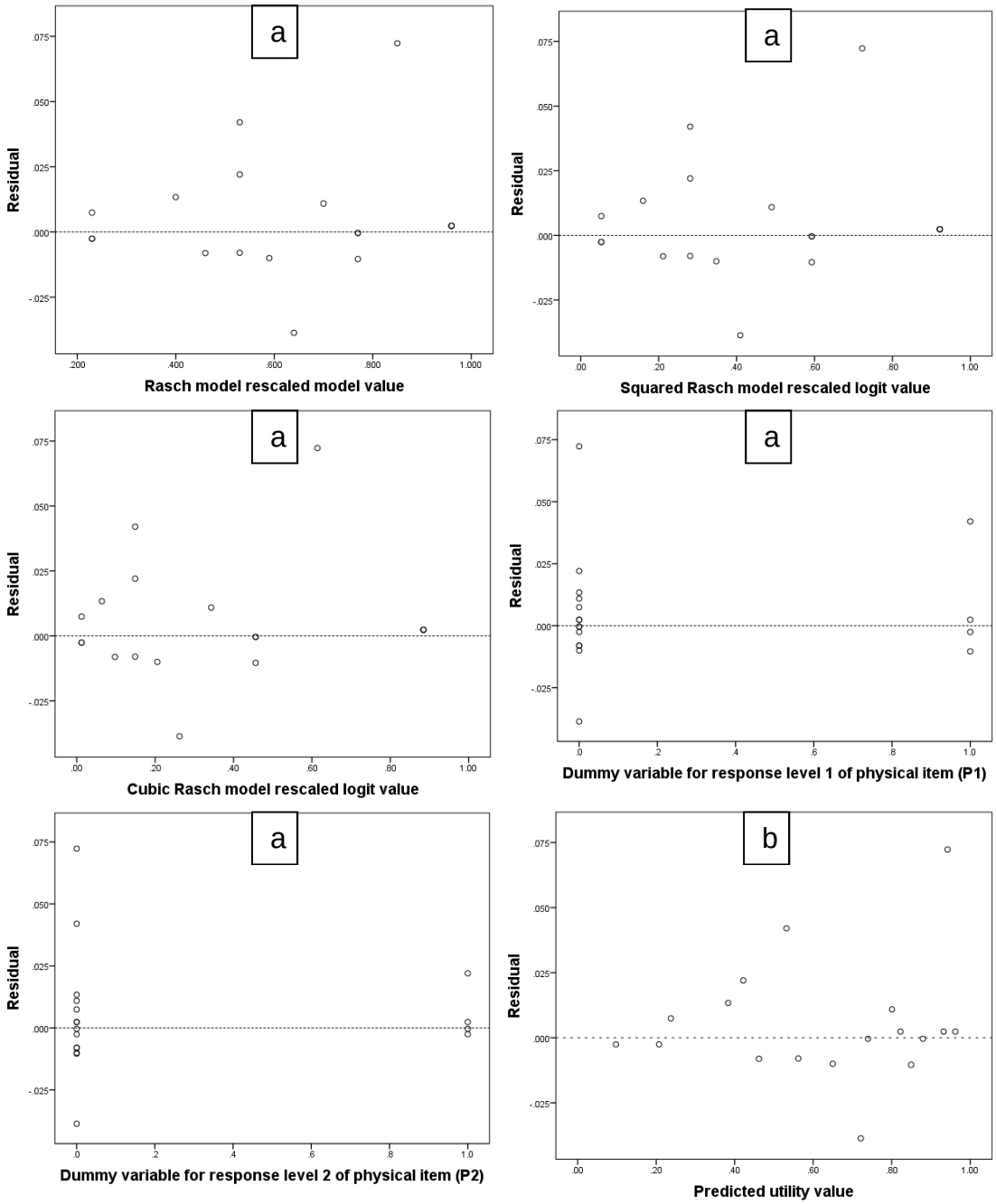
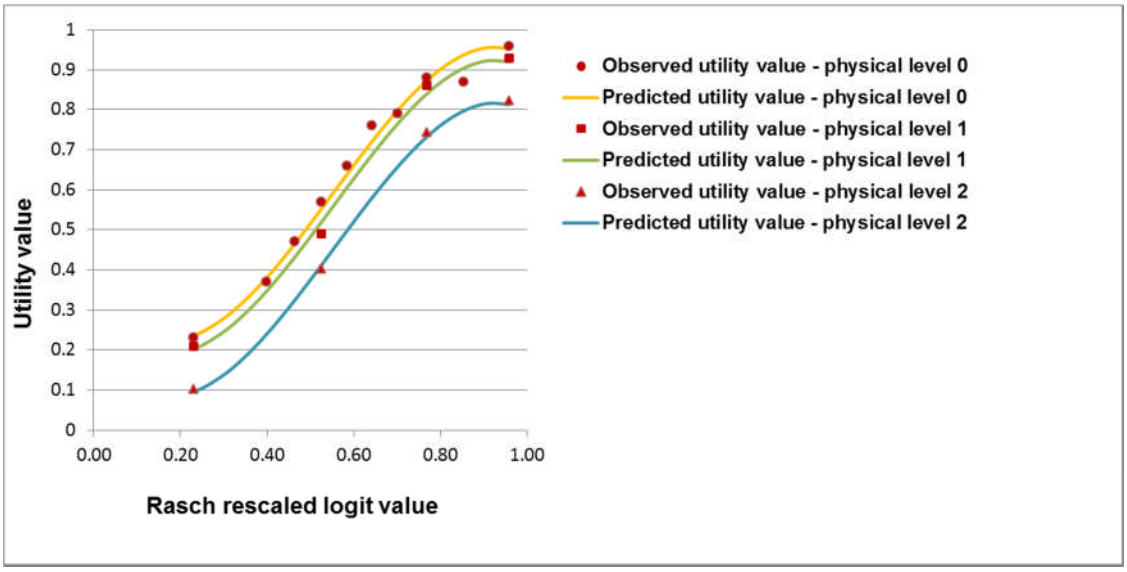


Figure 20 allows the comparison between actual mean utility values obtained from the valuation survey for the selected CORE-6D health states, and predicted utility values for all potential health states described by CORE-6D, derived from the mean-level base-case regression model m7. The x axis of the graph represents Rasch rescaled logit values that cover the full severity range of all potential emotional health states described by CORE-6D. The y axis depicts respective utility values. There are three lines on the graph, one for each level of the physical item. The 3 lines have an s-shape reflecting the cubic relationship between the Rasch logit scale and the health state utility value. Visual inspection of the plot of the observed and predicted utility data alongside the statistical performance of the base-case model m7 confirms that the model performs well.

Figure 20. Mean observed (from the valuation survey) and modelled (based on regression model m7) utility values by Rasch rescaled logit value



Note: Modelled utility values are predicted using the Rasch rescaled logit value of the emotional health state and the response level of the physical item 'I am troubled by aches, pains, physical problems' (level 0 = never; level 1 = only occasionally or sometimes; level 2 = often, most or all the time)

As already noted, the regression models described above can be used to estimate a utility value for every CORE-6D health state, based on the total (ordinal) score of its emotional component and the response level of the physical item. Table 38 reports the total (ordinal) emotional state score, the corresponding Rasch logit value (both original and rescaled), and the modelled utility values for all potential CORE-6D health states as estimated using the selected base-case regression model m7. It can be seen that the utility values of the CORE-6D index range from 0.10 (worst emotional and physical state) to 0.95 (best emotional and physical state). Utility values appear to be only mildly affected by moderate physical symptoms (i.e. response level 1), while severe physical symptoms (response level 2) seem to affect more substantially the estimated utility values. A syntax that allows calculation of CORE-6D utility values from CORE-OM data in SPSS is provided in Appendix 13.

Table 38. Modelled mean utility values for all CORE-6D health states, based on the total ordinal score of the emotional component of the state and the response level of the physical item, using the base-case regression model m7

CORE-6D			Response levels of the physical item		
Total emotional state score	**Corresponding original Rasch logit value**	**Corresponding rescaled Rasch logit value**	**0**	**1**	**2**
0	-3.748	0.96	0.95	0.92	0.81
1	-2.681	0.85	0.94	0.90	0.80
2	-1.836	0.77	0.87	0.84	0.73
3	-1.168	0.70	0.80	0.77	0.66
4	-0.573	0.64	0.72	0.69	0.58
5	0.002	0.59	0.64	0.61	0.50
6	0.594	0.53	0.55	0.52	0.41
7	1.217	0.46	0.47	0.43	0.32
8	1.878	0.40	0.38	0.35	0.24
9	2.652	0.32	0.30	0.26	0.16
10	3.562	0.23	0.24	0.20	0.10

Individual-level models

Individual-level regression analysis considered 4 different models:

Model i1 was an OLS model that used as explanatory variables the Rasch model rescaled logit value (linear, quadratic and cubic form) and 2 dummy variables for response levels 1 and 2 of the physical item of CORE-6D, that is, it used the same explanatory variables considered in the best-performing mean-level OLS base-case model m7 ('health state variables'). Model i1 did not take into account any socio-demographic characteristics.

Model i2 was an OLS model, which, in addition to the above explanatory health state variables, considered also socio-demographic variables with significant coefficients at the p=0.05 level, that is, age, gender and ethnicity. A preliminary OLS analysis that included a wider range of socio-demographic characteristics as explanatory variables revealed that a number of other variables, such as relationship status, home ownership, level of academic degree and employment status had non-significant coefficients; these variables were thus excluded from consideration in model i2. Age was entered in the regression as a continuous variable in linear and quadratic form (i.e. age-squared); gender and ethnicity were entered by introducing 2 dummy binary variables, female vs. male and white British background vs. any other ethnic background, respectively.

Model i3 was a Tobit model that considered the same explanatory variables with model i1 (that is, it included exclusively health state variables).

Model i4 was a Tobit model that contained the same explanatory variables with model i2 (that is, it considered both health state variables and significant socio-demographic variables).

Results of individual-level models are provided in Table 39. For each Tobit model, two pairs of results are provided: the regression coefficients which reflect how the unobserved, latent variable y^*_i changes with respect to changes in the independent variables, and the mean marginal effects of the

independent variables on the censored observed utility value y_i. It can be seen that the response level 1 of the physical item was non-significant in any of the models; the Rasch logit model value had significant regression coefficients at the 0.05 level only in Tobit models, whereas the quadradic and cubic forms of the Rasch logit model value as well as the response level 2 of the physical item had significant regression coefficients in all individual-level models. Both OLS and Tobit analyses indicated a statistically significant relationship between the observed utility values and age, gender and ethnicity (Table 39). Consideration of these variables improved to a small extent the model fit in both types of models (RMSE of individual-level OLS regression improved from 0.39 when no socio-demographic characteristics were considered, to 0.38, when socio-demographics were considered; the estimated standard error of the Tobit regression without socio-demographic characteristics was 0.51 and was reduced at 0.49, when socio-demographic variables were added).

Overall, inclusion of significant socio-demographic variables improved only marginally the fit of the models. More importantly, the model fit of the individual-level models was much lower than that observed for the mean-level models (Table 36), which reflects the large random variability at the individual level that is not needed for policy purposes, where mean-level models are more suitable to use.

An interesting finding of the individual-level analyses that included socio-demographic variables is the quadratic relationship between utility values and age, which represents an inverted U-shaped function. According to the marginal effects of the Tobit model (model i4), ceteris paribus, the preferences of a 24-year old person are likely to be the same with those of a 85-year old person; similarly, the utility values obtained by a 38-year old person should be the same with those elicited from a 71-year old person. The maximum utility value is obtained at 54 years of age.

Table 39. Results of individual-level least ordinal squares and Tobit regression models for the prediction of CORE-6D utility values

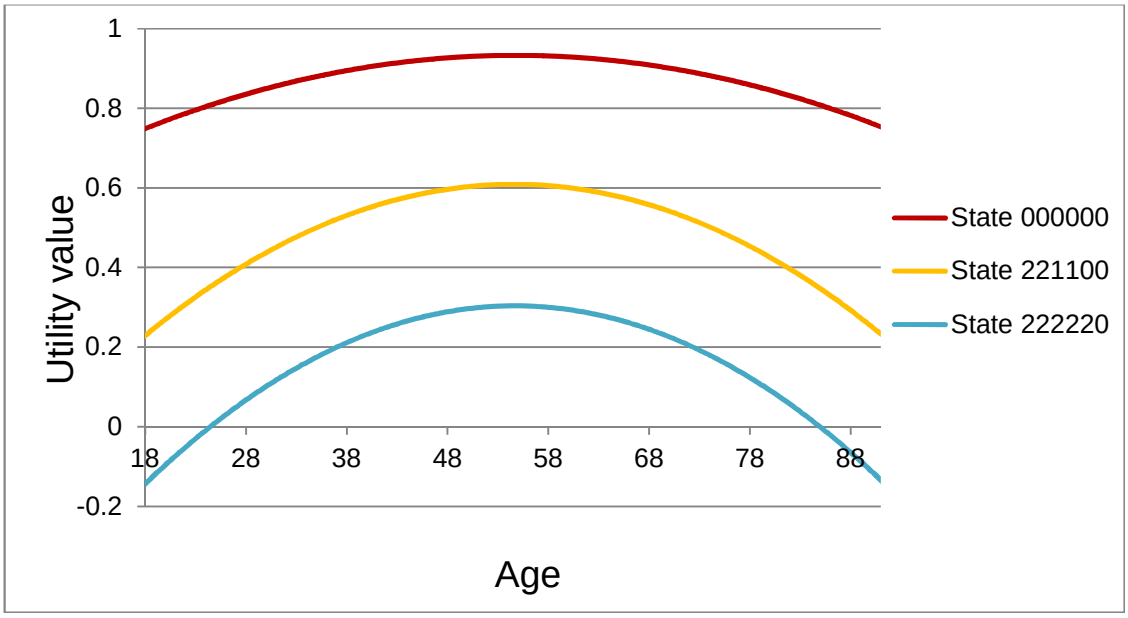
[analysis on N = 1,492 individual utility values]

Model	**(i1)**	**(i2)**	**(i3)**		**(i4)**	
Variable	**RC**	**RC**	**RC**	**ME**	**RC**	**ME**
Constant	0.375	-0.177	0.639	NA	-0.159	NA
	(0.035)	(0.351)	(0.009)		(0.533)	
Rasch model logit value of the emotional component (linear, quadratic and cubic form)						
R	-1.788	-1.904	-3.354	-2.399	-3.516	-2.546
	(0.125)	(0.095)	(0.036)	(0.036)	(0.024)	(0.023)
R^2	5.994	6.234	9.010	6.445	9.360	6.778
	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
R^3	-3.662	-3.819	-5.011	-3.584	-5.249	-3.801
	(0.003)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)
Dummy variables for response levels of the physical item other than zero						
P_1	-0.0232	-0.0161	-0.0383	-0.0277	-0.0285	-0.021
	(0.391)	(0.542)	(0.309)	(0.314)	(0.436)	(0.439)
P_2	-0.1281	-0.1264	-0.2032	-0.1529	-0.2010	-0.1531
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Dummy variables accounting for personal characteristics						
Age		0.0263			0.0379	0.0274
		(0.000)			(0.000)	(0.000)
Age squared		-0.0002			-0.0003	-0.0003
		(0.000)			(0.000)	(0.000)
*G*ender (female)		0.0651			0.0884	0.0645
		(0.002)			(0.002)	(0.002)
*E*thnicity (white British)		-0.1188			-0.1612	-0.1088
		(0.002)			(0.002)	(0.001)
Overall model stats						
Adjusted R-Squared	0.3249	0.3636	-		-	
RMSE/mean standard error	0.3949	0.3824	0.5131		0.4916	

Notes:
Model i1 – OLS regression, no socio-demographic characteristics considered
Model i2 – OLS regression including significant socio-demographic characteristics (p≤0.05)
Model i3 – Tobit regression, no socio-demographic characteristics considered
Model i4 – Tobit regression including significant socio-demographic characteristics (p≤0.05)
ME: mean marginal effects
RC: regression coefficients
RMSE: Root mean squared error
p values in parenthesis

The impact of age is also shown in Figure 21. The graph presents the predicted utility values against age, for men of white origin and for 3 different emotional health states, each combined with the physical item at response level 0: a. the mildest emotional state 00000 (producing the full state 000000); b. the intermediate emotional state 22110 (leading to full state 221100); and c. the worst emotional state 22222 (resulting in full state 222220). For all states, the relationship between utility values and age follows an inverted U-shaped function, with maximum utility value reached at 54 years of age. The shape becomes sharper as the emotional health states become more severe, indicating that the impact of age becomes more prominent as the emotional symptom severity increases.

Figure 21. Effect of age on predicted utility value (Model i4)



6.4 Discussion

This chapter describes the development of a utility index for the CORE-6D classification system, following a novel methodology that uses mainly the results of Rasch analysis. Rasch analysis was employed at the development of the unidimensional emotional component of CORE-6D, as described in detail in Chapter 5. Subsequently, it was employed for the identification and selection of plausible emotional health states that were considered in the valuation survey. Finally, Rasch analysis enabled prediction of utility values for all health states of CORE-6D by estimating the relationship between the Rasch model logit values and the mean observed utility values obtained from the valuation survey using regression analysis. This novel approach based on Rasch analysis was adopted because of the high correlations between the CORE-6D items that did not allow use of standard methods for generating and modelling health states.

Conventional approaches for generating health states, such as orthogonal block designs, treat items as independent (uncorrelated) statements (e.g. Brazier et al., 2002; Dolan et al., 1996). Therefore, such approaches are not appropriate for use in measures that have no clear multidimensional structure (that is, measures that consist of highly correlated items), such as the emotional component of CORE-6D, because they entail the danger of generating implausible health states (an example is a state that includes the statement “I have felt optimistic about my future” at the same time with the statement “I have made plans to end my life”). In contrast, consideration of the Rasch item threshold map for the identification of health states from CORE-6D helped to avoid generation of such implausible health states and, instead, create credible health states that comprised combinations of item responses observed in a real population.

An advantage of the ‘Rasch vignette approach’ is that it leads to the development of health states that 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 symptom severity based on his/her responses and vice versa. Indeed, the 11 emotional health states of CORE-6D

that were identified by inspection of the Rasch item threshold map were among the most frequently observed emotional health states in the study population: although these states represent only 4.5% of the potential 243 health states described by the emotional component of CORE-6D, they covered approximately 35% of the responses obtained from a large CORE-OM dataset that was analysed in this thesis [N1500]. In contrast, the coverage of the 15 health states generated from the emotional component of CORE-6D using an orthogonal block design reached only 14% in the same dataset; furthermore, use of an orthogonal block design generated a number of implausible health states.

The Rasch vignette approach for the identification of plausible health states can be compared with the clustering-based approach developed by Sugar and colleagues (1998), who, as described in Chapter 3 (section 3.3.2), used predominantly cluster analysis to group item responses and subsequently combined the most frequent individual item responses within each cluster in order to develop health state descriptions. However, in contrast to the Rasch vignette approach, the item response combinations derived from the clustering-based approach were not necessarily the most frequently observed in the study sample; what's more, it is possible that they had not been observed at all in the study sample.

One limitation of the Rasch vignette approach, similar to the methodology proposed by Sugar and colleagues (1998), 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 CORE-6D, the Rasch vignette approach led to identification of 11 plausible emotional health states, which, combined with 3 response levels of the 'physical' item of the original CORE-OM (I have been troubled by aches, pains, or physical problems), produce a 2-dimensional set of 33 plausible health states that were used as the basis for the valuation of the CORE-6D.

Nonetheless, despite generating a limited number of health states, the major advantage of the Rasch vignette approach over the clustering-based approach

is that it allows the valuation of all potential health states described by a unidimensional measure (such as the emotional component of CORE-6D) by assigning all potential health states (i.e. all combinations of item responses including those not illustrated in the Rasch item threshold maps) to different locations along the latent scale according to their level of severity. The relationship between the Rasch model logit values of the health states and the corresponding mean observed utility values obtained from a valuation survey can be then established using regression analysis and subsequently used to generate utility values for all people completing CORE-OM.

In the case of CORE-6D, the standard approach for modelling utility values, by creating dummy variables in regression analysis for each level of every item of the measure (Brazier et al., 2002; Dolan, 1997), would have required far more states to be valued, due to the high correlation between the items of the emotional component of CORE-6D. In contrast, Rasch analysis has proved to be a more efficient solution for modelling utility values in such cases. For this thesis, a mixed approach for modelling utility values was successfully developed, by combining the Rasch-based approach reported by Young and colleagues (2010) with the standard approach used to account for the different severity levels of the physical item of CORE-6D (Brazier et al., 2002; Dolan, 1997). A number of additive mean-level regression models were tested, which assumed that there is no utility interaction between the response level of the physical item and the severity level of the emotional component of CORE-6D.

The selected cubic model was characterised by a high adjusted R-Squared (0.990), low RMSE (0.0275), low MAE (0.014) and good predictive ability as indicated by low observed residuals [78% (14/18) of health states had a residual ≤ 0.01, 17% (3/18) had a residual higher than 0.01 but ≤ 0.05 and 6% (1/18) of the states had a residual ≤ 0.10]. Inspection of residual plots indicated no bias in the distribution of residuals around independent variables and around predicted utility values, suggesting a good model fit. Visual inspection of the plot of observed utility values against the predicted ones confirmed the good predictive ability of the model. These results compare favourably with

regression models described in similar modelling studies (Brazier et al., 2002 & 2008; Dolan, 1997; Yang et al., 2009 & 2011).

It needs to be noted that the number of independent variables fitted in the model (5) was relatively high compared with the small number of data points (18 mean-level utility values obtained in the valuation survey). The latter reflected the small number of health states that were included in the valuation survey, owing to funding restrictions. Although this is acknowledged as a limitation of the analysis, on the other hand, modelling at the mean rather the individual level of utility values was preferred, as prediction of mean utility values is more relevant in cost-utility analysis where the preferences of the population need to be taken into account (Brazier et al., 2007; Feeny et al., 2002). However, in cases where the number of independent variables is large relative to the number of data points, there is danger that the model is overfitting the data and the value of R-squared statistics is limited (Harrell, 2001). In such cases it is advised that only variables that are expected to be good predictors of the dependent variable (utility value) be included in OLS regression analysis (Harrell, 2001). The selected model m7 included 5 independent variables; 3 were forms of the Rasch logit model value (linear, quadratic and cubic) reflecting the level of emotional distress while 2 dummy variables expressed the level of physical impairment. Both emotional distress and physical impairment are conceptually expected to be good predictors of perceived HRQoL and hence of utility values. Four of the variables (all forms of Rasch logit model value and response level 2 of physical item) had statistically significant regression coefficients in the selected model, and one variable (response level 1 of the physical item) had a regression coefficient that was slightly above the 0.05 level of significance (p=0.07). Individual-level models showed that the quadratic and cubic forms of the Rasch logit model value and the response level 2 of the physical item had statistically significant regression coefficients in all respective analyses (both OLS and Tobit), while the linear form of the Rasch logit model value was significant in Tobit models. The response level 1 of the physical item was non-significant in any of the analyses either at the mean or the individual level; nevertheless, this variable was deemed to conceptually be affecting HRQoL and was thus retained in the

analysis. Other statistics indicated that the selected model m7 had good model fit and predictive ability. Therefore, this model was finally selected for the prediction of mean utility values from CORE-6D, although the limitations of this analysis are acknowledged.

Extra mean-level regression models that considered multiplicative interaction between the physical item and the emotional component of CORE-6D did not offer any improvement in the model fit compared with the selected model, thus suggesting that a simple additive model was adequate. This latter finding supports an assumption that the impact of different dimensions on preferences is additive. If the assumption holds, inclusion or exclusion of a dimension should lead to no significant change in the coefficients of the other dimensions in the classification. However, this was not found in another study where a pain dimension was added to an asthma-specific utility measure, the AQL-5D (Brazier et al., 2011). This resulted in the coefficients of 2 of the other dimensions being significantly changed. However, the case of AQL-5D is different from that of CORE-6D because the other dimensions of AQL-5D were primarily concerned with physical health and so were less independent from a pain dimension than the emotional component of CORE-6D from the measure's physical item.

Analysis of valuation data at an individual level showed that there was a small but significant relationship between the utility values obtained in the valuation survey and some of the socio-demographic characteristics of the respondents, including age, gender and ethnicity. This finding is broadly consistent with the results reported in the valuation of EQ-5D, where age, gender and marital status were found to significantly affect the utility values elicited in the valuation survey (Dolan, 2000; Dolan & Roberts, 2002). The influence of these variables on the utility values was weak to moderate, with marginal effects in the Tobit model ranging in magnitude from 0.03 (age) to 0.11 (ethnicity). Interestingly, individual-level analysis revealed an inverted U-shaped function between utility values and age (fitted with age and age-squared terms), which is comparable with the relationship between these two variables found at the valuation of EQ-5D (Dolan & Roberts, 2002) and SF-6D (Kharroubi et al.,

2007). These findings highlight the importance of eliciting preferences from a representative sample of the general population in terms of age, gender and ethnicity, since valuations elicited from non-representative samples may not reflect the general population's preferences.

Nevertheless, the analysis demonstrated that the Rasch model logit value of the emotional component of CORE-6D, which expresses the level of severity of emotional symptoms, was by far the most substantial determinant of utility values. In any event, analysis of valuation data at an aggregate (mean) level offered a better solution for prediction of utility values compared with individual level models. This is consistent with findings of previous research, according to which aggregate models may perform equally to or even better than individual-level ones, regardless of the presence of some significant socio-demographic factors, because they eliminate unhelpful individual-level variation (Brazier et al., 2002 & 2008; McKenna et al., 2008; Yang et al., 2009 & 2011).

The valuation of CORE-6D followed the MVH group TTO protocol that was developed for the valuation of EQ-5D (Dolan, 1997; Dolan et al., 1996). Adoption of this protocol including its visual props permitted comparability of CORE-6D with the EQ-5D and met previous NICE requirements according to which, when an alternative to EQ-5D is used, the same methods of valuation should be adopted (National Institute for Health and Clinical Excellence, 2008).

However, the MVH group TTO protocol suffers from a number of limitations. A major criticism involves the effect of respondents' age on valuations (Dolan, 2000; Dolan et al., 1996). It could be argued that framing the valuation statements using a 10-year time horizon, as in the MVH protocol, may feel too generous for older respondents and yet too short for younger ones. The time horizon used to frame the statements has indeed been shown to affect TTO valuations, with TTO utility values decreasing as the time horizon of the valuation statement increases (Lin et al., 2012; Stiggelbout et al., 1995). Further exploration of the impact of age on health state valuations suggests that differences in valuations between young and old respondents would have still been observed if respondents' life expectancy had been used rather than a

fixed time horizon of the valuation statements (Dolan & Roberts, 2002; Robinson et al., 1997).

A second major criticism of the MVH group TTO protocol relates to the procedure for the valuation of states that are worse than death. One problem of the procedure is that it includes the apparently unrealistic scenario of moving from poor health to full health. On the other hand, reversing the ordering of the states so that the move is from full health to poor health entails the danger that respondents may believe that they can commit suicide following the end of the period they spend in full health (Rowen & Brazier, 2011). Another problem of the TTO procedure for valuation of states worse than death is that it is different from the procedure for valuation of states better than death: the procedure for valuation of states worse than death assumes a total fixed duration of the time spent in full health plus the time spent in the health state subject to valuation, and varies the duration of the time spent in full health concurrently with the time in the health state subject to valuation, so that the total duration remains intact; in contrast, the procedure for the valuation of states better than death assumes a fixed duration for the period of time spent in the health state subject to valuation, and varies only the period spent in full health. Further to the confusion caused to the respondents, use of different procedures for valuation of states worse versus better than death creates a gap effect in the utility values for states around death (Tilling et al., 2010).

Finally the MVH protocol has been criticised for the monotonic transformation of values of states considered worse than death so that values are bounded by -1. As described in section 1.3.1 of Chapter 1, the need for this transformation was dictated by the very low values that were possible to obtain when the formula $-x/y$ was used, with implications at the modelling of valuation data. However, the transformation of values for states considered worse than death by using the formula $-x/10$ means that these values can no longer be interpreted as utility scores, and therefore are not comparable to utility values elicited for states better than death; consequently, aggregation of these two sets of values in econometric modelling becomes problematic (Patrick et al.,

1994). Possible resolutions to this problem comprise an alteration in the TTO protocol used to elicit values by introducing a 'lead time' in full health for both types of states (better and worse than death), and an alternative methodology in the modelling techniques used for the prediction of utility values, with both approaches being under on-going research (Rowen & Brazier, 2011).

It is acknowledged that the valuation of the CORE-6D suffers from the same limitations characterising the MVH protocol. As stated earlier, adoption of the same protocol was deemed necessary to ensure comparability of CORE-6D with EQ-5D, which was prerequisite for the use of CORE-6D in the NICE decision-making context. At the same time researchers have highlighted the importance of balancing between compliance with requirements of regulatory bodies and the use of currently best available methods (Feeny, 2013); any limitations in methodology advocated by such bodies should be identified, leading to prioritisation of areas for further research (Sculpher, 2013). It should be noted, though, that the most recently published NICE methods guidance no longer requires, at least explicitly, the same valuation method with that adopted for EQ-5D when alternative PBMs are used for the estimation of QALYs (National Institute for Health and Care Excellence, 2013).

6.5 Conclusion

The use of the novel methodology described in this chapter, based primarily on Rasch analysis, enabled the development of a utility index for CORE-6D, a 2-dimensional PBM that can be used for the estimation of QALYs in economic evaluations of interventions for common mental health problems. The resulting algorithm can be applied to any CORE-OM dataset prospectively or retrospectively. Application of the CORE-6D algorithm on existing CORE-OM datasets has been used to examine the performance of the new PBM. The results of this exercise are described in Chapter 7.

Chapter 7. Performance of CORE-6D: comparison with generic preference-based measures and the CORE-OM

7.1 Introduction

The purpose of developing CORE-6D was the formation of a new PBM that is specific to common mental health problems, an area where the use of generic PBMs appears to be rather limited, less acceptable and less responsive to HRQoL changes compared with other disease areas where physical symptoms prevail. However, before CORE-6D is widely used in economic evaluations for the estimation of QALYs, its psychometric properties and content need to be assessed in order to confirm that the new measure can capture appropriately any changes in the HRQoL of people with common mental health problems. This chapter presents the methodology and the results of a series of analyses that aimed to compare the CORE-OM with generic PBMs in terms of their psychometric properties, and also to explore the degree of loss of information resulting from the derivation of CORE-6D from the original CORE-OM. The methods adopted and the generic PBMs used as comparators for this purpose were dictated by the availability of relevant data in the area of common mental health problems.

7.2 Methods

7.2.1 Overview of compared measures

CORE-6D

The CORE-6D consists of an emotional component with 5 conceptual domains (symptoms - anxiety, functioning - general, functioning - close relationships, functioning - social relationships, risk /harm to self) and a physical health item. Each item has 3 levels of severity, which, combined, can produce 729 distinct health states; utility values range from 0.10 to 0.95.

SF-6D

The SF-6D can be derived from SF-36 (Brazier et al., 2002) as well as from its shortened version SF-12 (Brazier & Roberts, 2004). The SF-6D has 6 dimensions (physical functioning, role limitations, social functioning, bodily pain, mental health and vitality). When derived from SF-36, physical functioning and bodily pain have 6 levels of severity, role limitations 4 and the rest 3 dimensions have 5 levels of severity each; this form of SF-6D can describe approximately 18,000 unique health states and the resulting utility values range from 0.301 to 1. When derived from SF-12, SF-6D includes 3 levels of response for physical functioning, 4 levels of response for role limitations, and 5 levels of response for each of the remaining dimensions, resulting in the formation of roughly 7,500 unique health states. This version of SF-6D corresponds to utility values that range from 0.35 to 1.

EQ-5D

The EQ-5D has 5 dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety /depression). The version of EQ-5D used in the comparisons has 3 levels of severity ranging from ‘no problems’ to ‘severe problems’ (Brooks, 1996). The health state classification system therefore describes 243 unique health states; utility values for the UK population range from -0.59 to 1 (Dolan, 1997).

EQ-5D data were not available in 2 of the 3 datasets that were analysed. In these 2 datasets available SF-12 data were mapped onto EQ-5D utility values. As discussed in Chapter 1 (section 1.3.3), mapping refers to the estimation of a statistical relationship between two measures that allows prediction of values of one measure, which is not available in a dataset, using data from another measure that is included in the dataset. The relationship is determined by undertaking regression analysis on a separate dataset that has similar patient characteristics to the dataset of interest and contains both measures (Brazier et al., 2010).

Two studies reporting mapping algorithms between SF-12 and EQ-5D utility values were reviewed for this purpose; both used data from 12,967 adults

participating in a national US survey on medical expenditure in 2000 (details of which are available from http://www.meps.ahrq.gov/mepsweb/). The first algorithm was developed by Franks and colleagues (2004), who used OLS regression methods to map the physical and mental component summary scores of the SF-12 (PCS-12 and MCS-12, respectively) onto EQ-5D utility values for the UK population (Dolan, 1997). The methodology used in this study was rather crude, as it ignored individual responses to each of the items of SF-12. Moreover, use of OLS has theoretical limitations, given the skewed and bounded nature of the utility data.

The second published algorithm was developed by Gray and colleagues (2006), who applied multinomial logit regression and Monte Carlo simulation methods to generate predictions of EQ-5D responses using individual item responses and summary scores from the SF-12 as explanatory variables. The predicted EQ-5D responses were then linked to the UK EQ-5D tariff (Dolan, 1997). The approach was validated on data from 13,304 people aged above 16 years who participated in a national health survey conducted in England in 1996 (Department of Health et al., 1998). This more sophisticated approach made better use of the available SF-12 data and overcame some of the problems characterising the algorithm developed by Franks and colleagues (2004). Therefore it was chosen for the prediction of EQ-5D utility values from available SF-12 data. In order to obtain mapped utility values using this algorithm, 1,000 iterations of the probabilistic algorithm were run and the median values of the predicted 1,000 EQ-5D utility values were used in the analyses described in this chapter.

CORE-OM

The CORE-OM consists of 34 items, each with 5 levels of response, capturing 4 conceptual domains: subjective well-being, problems (depression, anxiety, physical symptoms, and trauma), functioning (general functioning, close relationships, social relationships), and risk (risk-to-self and risk-to-others). Depending on the level of response, each item is scored from 0 to 4. The CORE-OM clinical score is then calculated by adding all 34-item scores, multiplying by 10 and dividing by 34. The CORE-OM clinical score can get

values between 0-40, with 10 being considered the cut-off point between clinical and non-clinical cases. A clinical score 10 to <15 indicates mild psychological distress, 15 to <20 moderate distress, 20 to <25 moderate to severe distress, and 25 to 40 severe psychological distress (Barkham et al., 2006). A completed CORE-OM questionnaire is considered 'valid' if responses to no more than 3 items are missing (that is, at least 31 items have been completed). When up to 3 item scores are missing, the CORE-OM clinical score is calculated by adding the scores of all completed items, multiplying by 10 and dividing by the number of completed items (Evans et al., 2002).

7.2.2 Quantitative and qualitative analyses performed

A number of psychometric and statistical tests (described in detail in section 1.2.4 of Chapter 1) were undertaken to compare CORE-6D with generic PBMs and the CORE-OM; all statistical analyses were performed on SPSS 19 (IBM Corp., 2010). In addition, the content validity of CORE-6D was explored using qualitative assessment.

Acceptability

Acceptability of the measures to respondents was indirectly assessed by the percentage of missing data across all available observations.

Floor and ceiling effects

Floor and ceiling effects are observed when a large percentage of people in the sample are in the worst and best health state, respectively. The presence of such effects indicates that the instrument is not well targeted to the study population, as it cannot measure the whole range of health; consequently the instrument is unable to capture improvement (where there are ceiling effects) or deterioration (where there are floor effects) in health. Such effects have an impact on a measure's responsiveness to change over time. Floor and ceiling effects were estimated and reported across all available observations in each of the datasets used.

Responsiveness to change over time

Responsiveness expresses the ability of an instrument to capture known and important changes in the health of individuals, which may reflect therapeutic effects (Kirshner & Guyatt, 1985). To ensure comparability of the results, responsiveness of the measures was examined using SRM and ES. SRM is the mean change score of a measure between two different time points divided by the standard deviation of the change score; the ES is the mean change score of a measure between two time points divided by the standard deviation of the score at baseline. A value of SRM or ES around 0.2 to 0.3 has been deemed to indicate a small effect, a value around 0.5 a medium effect, and a value of 0.8 and above a large effect (Cohen, 1988). Statistical significance of differences was assessed using paired t-tests.

It should be noted that assessment of a measure's responsiveness with these criteria implies that there is a real improvement in health over time following treatment, which the measure is able to capture. However, treatment may in fact not be effective, at least not for all people in the study population, so a finding of low responsiveness of a measure does not necessarily mean that the measure cannot capture real changes in health over time – it may simply reflect the fact that no real health improvement occurred following treatment. In the datasets and mental health populations used in the analyses reported in this chapter there was no 'gold standard' measure that could verify that people's health indeed changed over time, but rather an *expectation* that health improved following treatment, which the PBMs attempted to capture. Therefore, the results of these analyses should be interpreted with caution.

Statistics based on SRM and ES were reported for observations containing available data for every measure of interest.

Prior to examination of SRM and ES, utility values generated from all PBMs were plotted by period to determine whether they showed comparable movements in HRQoL at different time points, thus indicating that the PBMs capture changes in HRQoL over time in a similar way.

Construct validity

Construct validity refers to the extent to which an instrument can measure an underlying 'construct' (Fitzpatrick et al., 1998 & 2007). It can be assessed by the ability of the measure to distinguish among groups that are known to differ in the underlying construct ('known groups validity'), or by quantitatively examining the correlations of the instrument with a set of other variables that have been designed to measure the same construct ('convergent validity').

The **known groups validity** of the measures was assessed by their ability to distinguish among groups that are known to differ in mental symptom severity and overall health. To allow comparison across measures, this was standardised by estimating the ES, which was calculated as the difference in mean scores between two adjacent groups of study participants with different levels of mental symptom severity or overall health, divided by the standard deviation of the scores obtained from the mildest of the two sub-groups. Magnitude of the ES was judged as previously described, with a value approximately 0.2 to 0.3 considered to indicate a small effect, a value around 0.5 a medium effect, and a value of 0.8 and above a large effect (Cohen, 1988). The statistical significance of differences in values between sub-groups (pairwise comparisons) was assessed using *t*-tests. The significance of the ability of the measures to distinguish across multiple distinct levels of mental symptom severity and overall health was assessed by analysis of variance (ANOVA), after applying a post-hoc F-test. Statistics were reported for observations containing available data for every measure of interest.

Prior to the assessment of known groups validity, utility values generated from all PBMs were plotted across sub-groups of study participants with different levels of mental symptom severity as well as different levels of overall health. Such plots allowed assessing whether utility values can measure an improvement in perceived HRQoL resulting from a clinical improvement in the condition of interest.

The convergence of PBMs with CSMs that capture changes in mental symptom severity was explored as an indicator of the **convergent validity** of

the PBMs. The extent of convergence was assessed by estimation of the Pearson correlation coefficients between each PBM and the relevant CSMs that were available in the datasets. Coefficients above 0.8 were considered strong; those around 0.5 were deemed moderate, and up to 0.3 were thought to be rather weak. The level of statistical significance of correlations was also estimated. Statistical tests were performed by making pairwise comparisons across all available observations.

Agreement between preference-based measures

PBMs could be seen as variables of the same class sharing the same metric and variance (i.e. they all generate utility values on the same 1-0 full health-death scale) but there are reasons for supposing they may be different, as they capture different dimensions of HRQoL. Therefore, the level of agreement across PBMs was assessed by estimating intraclass correlation coefficients (ICC) (McGraw & Wong, 1996). Strong agreements were deemed those with coefficients above 0.8; coefficients around 0.5 indicated moderate agreements; and agreements with coefficients up to 0.3 were deemed rather weak. The level of statistical significance of agreement was also estimated. Tests were undertaken by making pairwise comparisons across all available observations.

Differences in the content between CORE-6D items and items of generic preference-based measures

CORE-6D is by purpose more focused on mental health problems compared with generic PBMs. Nevertheless, despite the difference in focus, a number of CORE-6D items appear to be similar to generic PBM items. For example, the physical item of CORE-6D appears to be directly related to the ‘bodily pain’ item of SF-6D, and also to the ‘pain/discomfort’ item of EQ-5D. Both CORE-6D and SF-6D include a social functioning item. Also, SF-6D and EQ-5D have an explicit mental health item each. Such similar items across different PBMs are normally expected to capture the same underlying HRQoL dimension. In order to explore the extent of similarities (or differences) in the content of the different dimensions captured by CORE-6D and generic PBMs, Spearman rank correlations were estimated between each of the CORE-6D items, and each of the items of generic PBMs. Coefficients $\geq \left|0.40\right|$, which showed a

moderate to strong correlation, indicated that the items of CORE-6D and the generic PBM captured dimensions with similar aspects. The level of statistical significance of the correlation was also of interest. Correlations between items of CORE-6D and generic PBMs were examined in pairwise comparisons using all available observations.

Content validity

Qualitative assessment of the content validity of CORE-6D was attempted by comparing the content of the CORE-6D items with the content of the 7 major themes of HRQoL that were identified as most relevant to people with mental disorders by Brazier and colleagues (2014) and Connell and colleagues (2012), as described in section 2.3.3 of Chapter 2. The 7 major themes of HRQoL in people with mental disorders that were identified by the above studies are subjective well-being & ill-being; activity & functioning; social well-being, belonging & relationships; self-perception; control, autonomy & choice; hope & hopelessness; and physical health. For comparison, the assessment of the content validity of generic PBMs against these 7 themes of HRQoL, which was reported in section 2.3.3 of Chapter 2, is presented later in this chapter as well.

7.2.3 Datasets analysed

Data derived from three different UK studies were analysed: the first dataset included a sub-sample of participants in the adult psychiatric morbidity survey (PMS) conducted in Great Britain in 2000 (Singleton et al., 2001); the second dataset included people with mild to moderate anxiety and/or depression participating in a RCT of supervised self-help cognitive behavioural therapy (psychological health by assessing self-help education – PHASE programme) (Richards et al., 2003); and the third dataset consisted of data obtained from postnatal women recruited for a multicentre RCT assessing psychological interventions for postnatal depression (postnatal depression effectiveness randomised controlled trial - PoNDER) (Morrell et al., 2009a & 2009b).

PMS dataset

The PMS dataset contains data from a sub-sample of a much larger sample included in a national psychiatric morbidity survey that was conducted in the UK in 2000 (Singleton et al., 2001). All participants in this survey (8,580 adults) had completed the CIS-R, an interviewer-administered questionnaire that covers 14 non-psychotic symptoms. Total CIS-R score is an indication of the overall symptom severity: a score of ≥6 suggests symptoms of a mental disorder, a score of ≥12 indicates a significant level of symptoms, and a score of ≥18 denotes symptoms of a level likely to require treatment (Lewis et al., 1992). From the original survey sample, 3,536 respondents were selected for follow-up interviews approximately 18 months later (Singleton & Lewis, 2003). This follow-up sample included all participants in the initial survey who had scored ≥6 on the CIS-R, indicating some symptoms or the presence of a mental disorder, as well as a random sample of 20% of the survey respondents who had scored 0–5 on the CIS-R, indicating no mental disorder. After a second follow-up interview, participants were randomly allocated to complete one of three self-report paper measures of psychological wellbeing, with 682 individuals being allocated to complete the CORE–OM. Of these, 558 returned questionnaires, with 553 providing valid CORE-OM questionnaires (that is, questionnaires with at least 31 completed items). Data on these 553 respondents comprised the dataset used in this analysis. More details on the process of the selection of the 553 adults included in the PMS dataset analysed here are provided in Connell and colleagues (2007).

The dataset included responses to CORE-OM and SF-12, which allowed estimation of CORE-6D and SF-6D utility scores, respectively. SF-12 was also used in order to obtain mapped EQ-5D utility values. All data in the PMS dataset were collected at a single time point; no follow-up data were available.

PHASE dataset

The PHASE dataset consisted of adults participating in a RCT evaluating self-help cognitive behavioural therapy (CBT) facilitated by practice nurses against ordinary GP care (control group) for mild to moderate anxiety and/or depression, which was conducted in 17 general practices in north-east

England. The dataset included 112 study participants that had completed consent forms and the General Health Questionnaire-12 (GHQ-12), which was used to detect clinical cases (a mean score of 3 and above indicated a clinical case). Details on the selection of patients included in the dataset are reported in Richards and colleagues (2003).

The dataset included responses to CORE-OM and EQ-5D, which allowed estimation of CORE-6D and EQ-5D utility scores, respectively. Data were available at baseline, end of treatment, 1-month follow-up and 3-month follow-up (although there was no demarcated end of treatment for the control group, assessment occurred at the same time as end of treatment for the self-help CBT group to provide a 'matched' point of assessment). Due to large attrition rates, the 112 participants provided in total 214 observations across the 4 time points.

PoNDER dataset

The PoNDER dataset consisted of postnatal women recruited for a RCT assessing psychological interventions for postnatal depression, conducted in 101 general practices in Trent, England, between 2003 and 2006. The dataset included 3,689 women, of whom 3,437 provided data at baseline. Details on the selection of women for the study are reported in two publications (Morrell et al., 2009a & 2009b). Women diagnosed with postnatal depression according to an Edinburgh postnatal depression scale (EPDS) score ≥ 12 were offered treatment (either by trained health visitors or standard care).

The dataset included responses to CORE-OM and SF-36, which allowed estimation of CORE-6D and SF-6D utility scores, respectively. SF-36 was also used to derive SF-12 data and subsequently mapped EQ-5D utility values. Data were available at baseline (6 weeks postnatally) as well as at 6-month, 12-month, and 18-month follow-up. Analyses were undertaken on the whole sample of 3,689 participants, who provided 9,439 observations across the 4 time points. In addition to these analyses, estimation of SRM and ES were also performed on a sub-sample of women that were diagnosed with postnatal

depression at baseline and had no missing data of interest at any time point in the study.

7.2.4 Determining known groups of different mental symptom severity and overall health in the datasets

The CORE-OM clinical score was used to determine different levels of mental symptom severity across observations in all datasets, according to the severity categories described earlier, due to lack of availability of other relevant measures in the datasets. In the PMS dataset, CIS-R data were also used to stratify respondents to different mental symptom severity levels: a score of 0-5 indicates little evidence of a mental disorder; 6-11 suggests some symptoms of mental disorder; ≥12 indicates symptoms at a clinical level; and a score of 18-63 denotes symptoms of a level likely to require treatment (Lewis et al., 1992). The level of overall health of each respondent in the PMS and PoNDER datasets was determined using their responses to the general health item 1 of the SF-12 ("in general, would you say your health is 1. excellent 2. very good 3. good 4. fair 5. poor"). The latter allowed assessment of the ability of CORE-6D to distinguish across different levels of overall (physical and mental) health, which is nonetheless relevant to a population with mental health problems and thus a desirable property for a mental health-specific PBM.

7.3 Results

7.3.1 Study sample characteristics and descriptive statistics

Sample characteristics

The PMS study sample had a mean age of 44.3 years (standard deviation 14.4); 43% of the sample were male. The mean CORE-OM clinical score in the study sample was 6.36 (standard deviation 5.19). Of the 553 respondents, 115 comprised clinical cases according to a CORE-OM clinical score ≥10 (20.8% of the sample). The mean CIS-R score was 7.40 (standard deviation 7.86); 122 respondents comprised clinical cases according to a CIS-R score ≥12 (22.1% of the sample). In the PHASE dataset, the mean age of the 112 participants in the trial was 39.3 years (standard deviation 12.7); 23.3% of the participants were male. The mean CORE-OM clinical score at baseline was 19.55

(standard deviation 5.22). Of the 106 participants who provided CORE-OM clinical scores at baseline, 103 (97.2%) scored at or above the clinical cut-off level of 10. Data were available for 109 people at baseline and 54 people at the end of treatment, while at the 1-month and 3-month follow-ups only 33 and 18 observations, respectively, were available. Finally, the mean age of the PoNDER sample was 31.1 years (standard deviation 5.4). Of the 3,437 women that provided responses at baseline (out of the 3,689 that were included in the dataset), 595 (17.3%) were diagnosed with postnatal depression according to an EPDS score ≥ 12, and 602 (17.5%) had a CORE-OM clinical score ≥ 10.

Descriptive statistics

As illustrated in Table 40, all PBMs generated a wide range of utility values in all 3 datasets, approximating each measure's utility range, with the exception of CORE-6D in the PMS dataset, where the range of its values was somewhat narrower compared with the other two datasets (range 0.71 in the PMS dataset vs. 0.85 in the PHASE and PoNDER datasets). The range of SF-6D values was the narrowest, approximately 0.65 in both PMS and PoNDER datasets, while EQ-5D (both direct and mapped) covered the widest range of values (range ≥1 in all datasets), reflecting the (inherent) wider utility range of EQ-5D compared with the other two PBMs.

Overall, the mean value of CORE-6D was higher than that of the generic PBMs, while its standard deviation was lower; the measure with the next highest mean value was the mapped EQ-5D. Mean utility values of all PBMs were higher in the PMS and PoNDER datasets compared with the PHASE dataset, likely reflecting the high percentage of non-clinical cases in the PMS and PoNDER study samples (approximately 80% in each dataset at baseline).

Table 40. Descriptive statistics, acceptability and floor and ceiling effects of all measures across the 3 datasets examined

Dataset	Measure	% missing data	Mean (SD)	Minimum	Maximum	% at floor (worst state)	% at ceiling (best state)	% at ceiling (best state) in clinical cases (CORE-OM ≥10)	
PMS (n=553)	CORE-6D	2.0	0.86 (0.11)	0.24	0.95	0.0	22.3	3.5	(n=115)
	SF-6D	0.9	0.77 (0.14)	0.37	1.00	0.0	1.5	0.0	
	Mapped EQ-5D	0.0	0.80 (0.23)	-0.14	1.00	0.0	36.0	5.2	
	CORE-OM	0.0	6.36 (5.19)	0.00	30.00	0.0	8.6	NA	
PHASE - all observations (n=214)	CORE-6D	9.7	0.70 (0.18)	0.10	0.95	0.5	4.6	0.0	(n=161)
	EQ-5D	6.9	0.60 (0.32)	-0.18	1.00	0.0	9.4	1.3	
	CORE-OM	5.5	16.27 (7.73)	0.00	37.65	0.0	1.0	NA	
PoNDER - all observations (n=9,439)	CORE-6D	1.1	0.89 (0.09)	0.10	0.95	0.0	34.6	0.4	(n=1,391)
	SF-6D	3.1	0.77 (0.12)	0.34	1.00	0.0	1.4	0.1	
	Mapped EQ-5D	0.0	0.86 (0.16)	-0.18	1.00	0.0	43.6	3.5	
	CORE-OM	0.6	5.04 (5.10)	0.00	36.47	0.0	19.0	NA	

CORE-OM scores were considered to correspond to full health if they had a value between 0-0.99; SD = standard deviation

7.3.2 Results of the psychometric analyses and statistical tests

Acceptability

The percentage of missing values for CORE-6D were within the range of 1-2% in the PMS and PoNDER datasets, which overall compared favourably with the respective figures of the other PBMs (Table 40); no missing values were recorded for mapped EQ-5D due to the probabilistic nature of the algorithm – so this result does not reflect the measure's acceptability. The CORE-OM demonstrated lower percentages of missing data, as CORE-OM scores can be estimated even when up to 3 item responses are missing. In the PHASE dataset the percentage of missing values was higher for CORE-6D (9.7%) compared with EQ-5D (6.9%), and both values were higher than the respective figure for CORE-OM (4.2%).

Floor and ceiling effects

No PBM showed floor effects (Table 40); yet mapped EQ-5D and, in a somewhat lesser degree, CORE-6D suffered from ceiling effects in the PMS (36.0% and 22.3%, respectively) and the PoNDER (43.6% and 34.6%, respectively) datasets, providing a first indication that they may not be able to cover the whole severity range in the study population. However, the majority of cases included in both these datasets (roughly 80% in each dataset) were non-clinical cases according to their CORE-OM clinical score, and therefore the high proportion of responses stating full health is not surprising. The presence of ceiling effects in these datasets was subsequently investigated following exclusion of all non-clinical cases, as defined by a CORE-OM clinical score <10, from analysis. Ceiling effects were reduced to a large extent, falling at 3.5% for CORE-6D and 5.2% for mapped EQ-5D in the PMS dataset, and 0.4% for CORE-6D and 3.5% for mapped EQ-5D in the PoNDER dataset. SF-6D showed a minimal ceiling effect in both PMS and PoNDER full datasets (around 1.4%). In the PHASE dataset, the percentage of ceiling effects for EQ-5D and CORE-6D was 9.4% and 4.6%, respectively, and were eliminated when only clinical observations (CORE-OM clinical score ≥10) were considered.

Responsiveness to change over time

Changes in the values of all PBMs over time in the PoNDER dataset are shown in Figure 22. Figure 22a shows data for cases with full utility data available across all time points (n=716), while Figure 22b shows cases with a diagnosis of postnatal depression (based on EPDS ≥12) at baseline *and* with full utility data available across all time points (n=103). All 3 measures detected comparable changes in HRQoL over time, indicating that the measures capture changes over time in a similar way. CORE-6D utility values were always higher than those generated using generic PBMs; SF-6D had the lowest values at each time point. Changes in utility values beyond the time point of 6 months were very small for all PBMs, both in the whole study sample and in the sub-sample of women with postnatal depression at baseline. Utility values generated from PBMs included in the PHASE dataset were not plotted by period, due to the small number of observations with data on both CORE-6D and EQ-5D (n=28 at 1-month follow-up and n=14 at 3-month follow-up).

Figure 22. Utility values plotted by period, PoNDER dataset

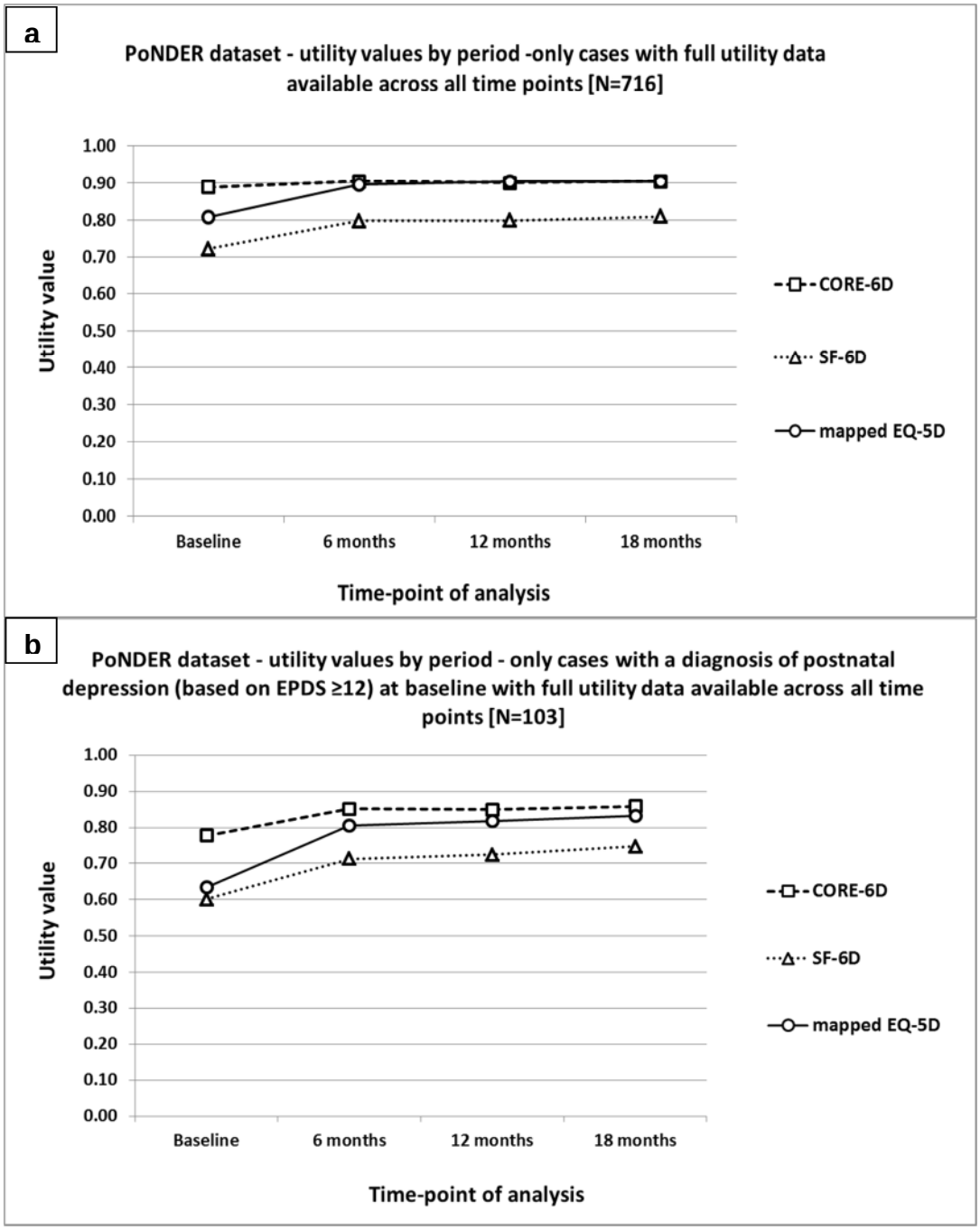


Table 41 provides findings on the mean change (and standard deviation), SRM and ES of each measure between baseline and different time points of analysis in the PHASE and PoNDER datasets. The table also presents the *p*-values of the paired *t*-tests that assessed the significance of change over time for each measure. In the PHASE dataset, CORE-6D showed a higher SRM and ES than EQ-5D between baseline and each time point examined (i.e. end of treatment, 1-month follow-up, and 3-month follow-up). CORE-6D demonstrated a moderate to large SRM and ES across different time periods, whereas the EQ-5D was characterised by negligible to moderate SRM and ES. The CORE-OM showed consistently higher responsiveness than CORE-6D, with values of SRM and ES exceeding 0.8 over any time period examined. Paired t-tests verified the significance of the results at the 0.05 level, with the exception of results obtained for the two PBMs between baseline and 3-month follow-up, which lacked statistical significance, possibly due to the very low number of observations (n=10). The number of observations used to compare responsiveness of the measures in the PHASE dataset was in general very small, and therefore the results should be interpreted with caution.

Results in the PoNDER dataset indicate that CORE-6D and CORE-OM had minimal (but statistically significant) responsiveness to change over time, with SRM and ES for both measures lying below 0.2 between baseline and the 3 time points examined (i.e. 6 months, 12 months, and 18 months). In contrast, SF-6D and mapped-EQ-5D showed a moderate to large (and significant) effect, with respective values of SRM and ES ranging from 0.51 to 0.75. In the sub-analysis that included only women diagnosed with postnatal depression at baseline, generic PBMs and the CORE-OM reflected a large, significant effect across all time points (all SRM and ES were above 0.75). CORE-6D showed a somewhat smaller but also significant responsiveness, with SRM and ES values ranging from 0.54 to 0.66, probably due to some loss of information relative to the CORE-OM.

Table 41. Responsiveness to change over time: standardised response mean and effect size

	PHASE dataset – only participants with data available for all 3 measures											
Measure	**Baseline to end of treatment [N=39]**				**Baseline to 1-month follow-up [N=21]**				**Baseline to 3-month follow-up [N=10]**			
	Mean change (SD)	**SRM**	**ES**	***t*-test p value**	**Mean change (SD)**	**SRM**	**ES**	***t*-test p value**	**Mean change (SD)**	**SRM**	**ES**	***t*-test p value**
CORE-6D	0.09 (0.19)	0.45	0.49	0.008	0.16 (0.21)	0.78	0.85	0.002	0.12 (0.23)	0.52	0.54	0.134
EQ-5D	0.10 (0.27)	0.38	0.35	0.021	0.12 (0.18)	0.68	0.44	0.005	0.00 (0.27)	0.01	0.01	0.973
CORE-OM	-6.27 (7.48)	-0.84	-1.04	<0.001	-7.77 (7.99)	-0.97	-1.19	<0.001	-7.32 (8.90)	-0.82	-0.90	0.029
	PoNDER dataset – only women with data available across all time points [N=716]											
Measure	**Baseline to 6 months**				**Baseline to 12 months**				**Baseline to 18 months**			
	Mean change (SD)	**SRM**	**ES**	***t*-test p value**	**Mean change (SD)**	**SRM**	**ES**	***t*-test p value**	**Mean change (SD)**	**SRM**	**ES**	***t*-test p value**
CORE-6D	0.02 (0.08)	0.19	0.18	<0.001	0.02 (0.09)	0.18	0.20	<0.001	0.02 (0.08)	0.19	0.20	<0.001
SF-6D	0.08 (0.12)	0.63	0.65	<0.001	0.08 (0.13)	0.58	0.66	<0.001	0.09 (0.13)	0.65	0.75	<0.001
Mapped EQ-5D	0.09 (0.17)	0.52	0.53	<0.001	0.10 (0.18)	0.53	0.58	<0.001	0.10 (0.19)	0.53	0.59	<0.001
CORE-OM	-0.70 (4.27)	-0.16	-0.14	<0.001	-0.65 (4.59)	-0.14	-0.14	0.001	-0.46 (4.67)	-0.10	-0.10	0.009
	PoNDER dataset – only women with postnatal depression at baseline, with data available across all time points [N=103]											
Measure	**Baseline to 6 months**				**Baseline to 12 months**				**Baseline to 18 months**			
	Mean change (SD)	**SRM**	**ES**	***t*-test p value**	**Mean change (SD)**	**SRM**	**ES**	***t*-test p value**	**Mean change (SD)**	**SRM**	**ES**	***t*-test p value**
CORE-6D	0.07 (0.13)	0.57	0.61	<0.001	0.07 (0.14)	0.54	0.61	<0.001	0.08 (0.13)	0.64	0.66	<0.001
SF-6D	0.11 (0.12)	0.87	1.48	<0.001	0.12 (0.12)	1.01	1.62	<0.001	0.14 (0.14)	1.03	1.93	<0.001
Mapped EQ-5D	0.17 (0.21)	0.81	0.93	<0.001	0.18 (0.24)	0.77	0.99	<0.001	0.20 (0.24)	0.82	1.08	<0.001
CORE-OM	-5.45 (6.08)	-0.90	-1.10	<0.001	-5.65 (6.74)	-0.84	-1.14	<0.001	-5.45 (7.06)	-0.77	-1.10	<0.001

ES = effect size; SD = standard deviation; SRM = Standardised Response Mean; *t*-test = *paired t*-test

Construct validity

Known groups validity

Utility values were first plotted across groups of different symptom severity (Figure 23) and different levels of overall health (Figure 24) in the PHASE and PoNDER datasets. Utility values across different CORE-OM clinical scores were not plotted for the PMS dataset, due to the very low number of people assigned to the more severe categories (n=38 in the 3 more severe categories combined). The graphs demonstrated that utility values increased in a consistent way as mental symptom severity decreased and general health increased. Overall, CORE-6D demonstrated the highest utility values among all PBMs and a shallower gradient compared with direct and mapped EQ-5D utility values. SF-6D had the narrowest range of values among PBMs between the worst and the mildest mental symptom severity groups as determined by the CORE-OM clinical score. When symptom severity was determined by CIS-R (PMS dataset), the range of values of all PBMs was narrower compared with the utility range under CORE-OM-defined symptom severity. CORE-6D had consistently the narrowest range of utility values between the worst and best general health level, as defined by responses to question 1 of the SF-12.

Figure 23. Utility values plotted by level of symptom severity

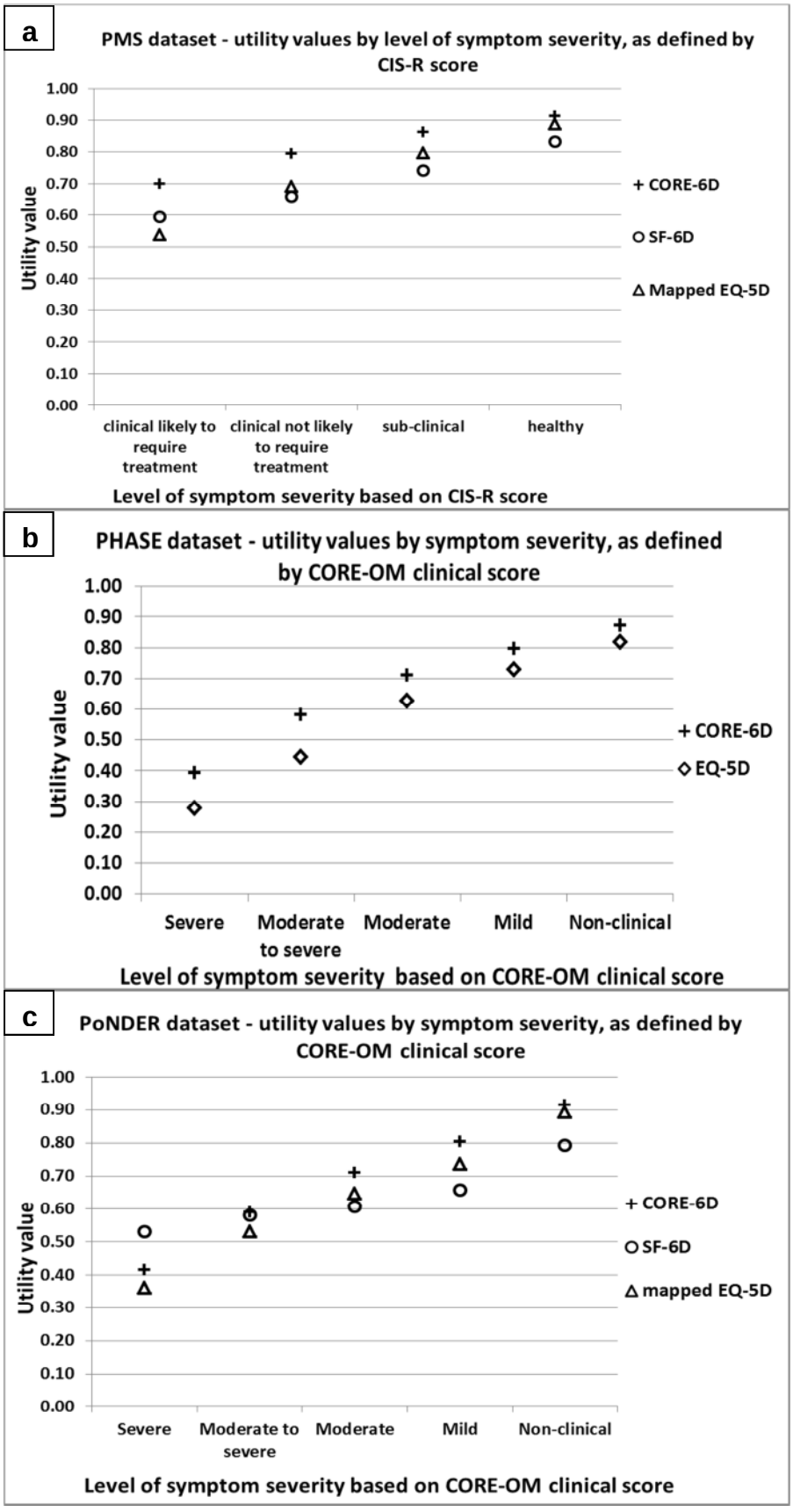
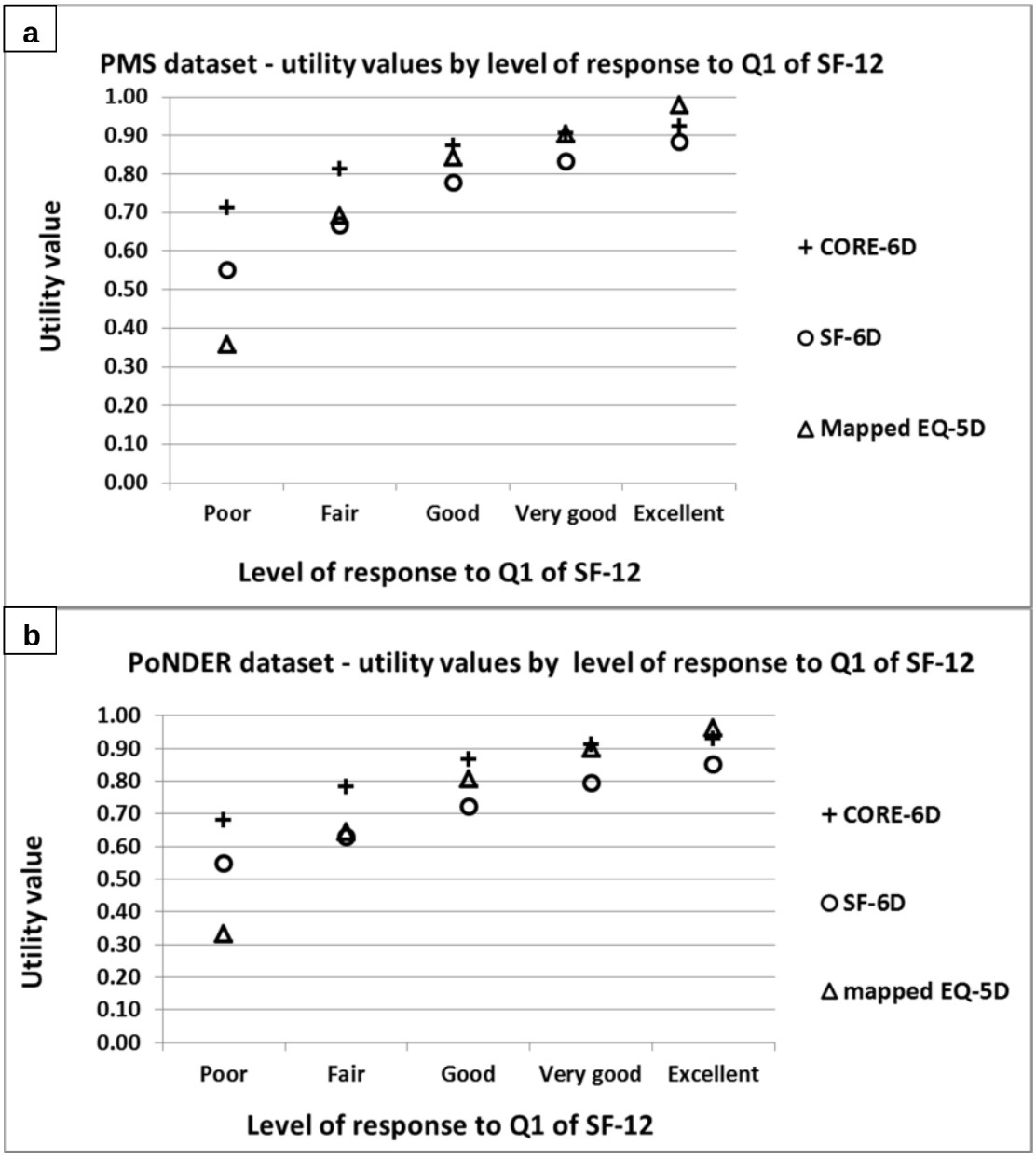


Figure 24. Utility values plotted by level of general health



CORE-6D and CORE-OM demonstrated the highest ES across different mental symptom severity groups, defined by either CORE-OM (Table 42) or CIS-R scores (Table 43). The ES values of CORE-6D exceeded 1 across all adjacent CORE-OM symptom severity groups in all datasets, and reached 1 when symptom severity was determined by CIS-R. The ES values observed for CORE-OM were substantially higher when known groups were defined by CORE-OM scores, ranging from 2.67 to 4.38 (Table 42). When known groups were defined by CIS-R, the CORE-OM ES was within the range of 1.0-1.5 (Table 43). T-tests confirmed the statistical significance of differences in the CORE-6D values and CORE-OM scores in pairwise comparisons between all different symptom severity groups ($p<0.01$).

In contrast, generic PBMs failed to discriminate between some levels of adjacent symptom severity determined by the CORE-OM clinical score, especially in the more severe end of symptoms, in the PMS and PHASE datasets. However, lack of statistical significance is possibly attributable to the small numbers of observations in the more severe end of symptoms in both PMS and PHASE datasets. In the PoNDER dataset, mapped EQ-5D was able to distinguish across adjacent levels of symptom severity, while SF-6D failed to distinguish between more severe symptom levels. Both SF-6D and mapped EQ-5D were able to distinguish across different CIS-R severity groups in the PMS dataset. The ES values of generic PBMs were in most cases considerably lower than those of CORE-6D (the ES of which was in the range of 0.79-1.92), and varied between 0.36-1.22 for the SF-6D (except between the two most severe CORE-OM levels in PMS where SF-6D moved in the opposite direction from that expected), 0.39-0.60 for EQ-5D and 0.06-1.22 for mapped EQ-5D, with higher ES values found in milder levels of symptom severity (Table 42 and Table 43). The ES values of all PBMs were comparable in the adjacent milder levels of symptom severity in the PoNDER dataset. ANOVA confirmed the statistical significance of differences of all measures across all groups of different symptom severity, determined by either the CORE-OM or the CIS-R clinical score.

Table 42. Known groups validity for different levels of mental symptom severity determined by the CORE-OM clinical score

		Severe	Moderate to severe	Moderate	Mild	Non-clinical	ANOVA: <0.001[1]
PMS dataset			[N=11][2]	[N=27]	[N=71]	[N=425]	T-tests[3]
CORE-6D	**Mean (SD)**		0.52 (0.16)	0.65 (0.10)	0.79 (0.09)	0.90 (0.06)	<0.001
	ES		1.30	1.46	1.92		
SF-6D	**Mean (SD)**		0.56 (1.12)	0.55 (0.09)	0.66 (0.12)	0.80 (0.12)	<0.001 to 0.055 1.000[4]
	ES		-0.11	0.93	1.15		
Mapped EQ-5D	**Mean (SD)**		0.45 (0.33)	0.47 (0.31)	0.70 (0.23)	0.86 (0.18)	<0.001 to 0.001 1.000[4]
	ES		0.06	1.04	0.89		
CORE-OM	**Mean (SD)**		24.16 (3.20)	17.42 (1.54)	11.75 (1.35)	4.19 (2.58)	<0.001
	ES		-4.38	-4.20	-2.93		
PHASE dataset		**[N=26]**	**[N=47]**	**[N=49]**	**[N=39]**	**[N=44]**	
CORE-6D	**Mean (SD)**	0.39 (0.14)	0.58 (0.12)	0.71 (0.10)	0.80 (0.08)	0.87 (0.07)	<0.001 to 0.008
	ES	1.57	1.30	1.08	1.06		
EQ-5D	**Mean (SD)**	0.28 (0.29)	0.44 (0.31)	0.63 (0.31)	0.73 (0.17)	0.82 (0.23)	<0.001 to 0.015 >0.05[5]
	ES	0.53	0.60	0.59	0.39		
CORE-OM	**Mean (SD)**	27.88 (2.75)	22.41 (1.58)	17.14 (1.38)	12.77 (1.32)	5.00 (2.91)	<0.001
	ES	-3.46	-3.82	-3.31	-2.67		
PoNDER dataset		**[N=39]**	**[N=124]**	**[N=385]**	**[N=799]**	**[N=7,577]**	
CORE-6D	**Mean (SD)**	0.42 (0.14)	0.59 (0.11)	0.71 (0.10)	0.80 (0.09)	0.92 (0.05)	<0.001
	ES	1.61	1.22	1.10	1.29		
SF-6D	**Mean (SD)**	0.53 (0.08)	0.58 (0.08)	0.61 (0.07)	0.66 (0.09)	0.79 (0.11)	<0.001 >0.100[6]
	ES	0.61	0.36	0.58	1.22		
Mapped EQ-5D	**Mean (SD)**	0.36 (0.22)	0.53 (0.24)	0.65 (0.18)	0.74 (0.16)	0.89 (0.13)	<0.001
	ES	0.70	0.63	0.56	1.22		
CORE-OM	**Mean (SD)**	27.79 (2.77)	22.00 (1.44)	17.02 (1.43)	12.03 (1.45)	3.23 (2.58)	<0.001
	ES	-4.02	-3.48	-3.44	-3.41		

ANOVA = analysis of variance; ES = effect size; SD = standard deviation

1. Significant for all measures in all datasets across different levels of symptom severity defined by CORE-OM
2. Severe & moderate to severe levels were merged due to small number of observations
3. Level of significance in pairwise comparisons between different severity levels
4. Between 'moderate' and the merged level 'moderate to severe and severe'
5. Between all adjacent symptom severity levels except between 'moderate' and 'moderate to severe'
6. Between 'moderate' and 'moderate to severe' as well as between 'moderate to severe' and 'severe'

Table 43. Known groups validity for different levels of symptom severity determined by the CIS-R score - PMS dataset

Symptom severity level – by CIS-R score						ANOVA: <0.001[1]
Measure		**18+ [N=58]**	**12-17 [N=62]**	**6-11 [N=122]**	**0-5 [N=292]**	**T-tests[2]**
CORE-6D	**Mean (SD)**	0.70 (0.16)	0.79 (0.10)	0.86 (0.09)	0.91 (0.05)	<0.001
	ES		0.94	0.79	1.00	
SF-6D	**Mean (SD)**	0.59 (0.11)	0.66 (0.13)	0.74 (0.12)	0.83 (0.11)	<0.001 to 0.009
	ES		0.52	0.67	0.87	
Mapped EQ-5D	**Mean (SD)**	0.54 (0.31)	0.69 (0.26)	0.80 (0.18)	0.89 (0.15)	<0.001 to 0.003
	ES		0.60	0.60	0.60	
CORE-OM	**Mean (SD)**	15.34 (5.94)	10.24 (3.47)	6.34 (3.26)	3.61 (2.69)	<0.001
	ES		-1.47	-1.20	-1.01	

ANOVA = analysis of variance; ES = effect size; SD = standard deviation
1. Significant for all measures across different levels of symptom severity defined by CIS-R
2. Level of significance in pairwise comparisons between different severity levels

As shown in Table 44, both CORE-6D and CORE-OM were able to distinguish between all adjacent general health levels, determined by responses to question 1 of SF-12, in both PMS and PoNDER datasets, with the exception of 'excellent' and 'very good' general health in the PMS dataset (pairwise t-tests non-significant). The ES of CORE-6D was higher than that of CORE-OM at the worst end of health. The ES of CORE-6D ranged between 0.31 and 0.91. The ES of CORE-OM had a narrower range, from 0.45 to 0.80. Generic PBMs were also able to distinguish across different levels of general health and overall demonstrated higher ES compared with the CSMs (especially the EQ-5D), a finding that was expected given the nature of question 1 of SF-12 as a general health item that can be deemed to mostly represent physical health. The overall range of ES values in both PMS and PoNDER datasets was 0.56-0.96 for SF-6D and 0.61-1.57 for mapped EQ-5D. ANOVA established the statistical significance of differences of all measures across all groups of different levels of general health, as determined by responses to question 1 of SF-12.

Table 44. Known groups validity for different levels of general health determined by responses to question 1 of the SF-12

		Poor	Fair	Good	Very good	Excellent	ANOVA: <0.001[1]
PMS dataset		**[N=46]**	**[N=100]**	**[N=172]**	**[N=148]**	**[N=68]**	**T-tests[2]**
CORE-6D	**Mean (SD)**	0.71 (0.15)	0.81 (0.11)	0.87 (0.09)	0.91 (0.06)	0.92 (0.05)	<0.001 to 0.009
	ES		0.91	0.70	0.57	0.31	1.000[3]
SF-6D	**Mean (SD)**	0.55 (0.107)	0.67 (0.14)	0.78 (0.11)	0.83 (0.09)	0.88 (0.06)	<0.001 to 0.013
	ES		0.86	0.96	0.65	0.86	
Mapped EQ-5D	**Mean (SD)**	0.36 (0.29)	0.69 (0.21)	0.84 (0.13)	0.90 (0.10)	0.98 (0.06)	<0.001 to 0.007
	ES		1.57	1.17	0.61	1.38	
CORE-OM	**Mean (SD)**	12.54 (6.59)	8.72 (5.33)	5.98 (4.49)	4.44 (3.46)	3.21 (2.50)	<0.001 to 0.020
	ES		-0.72	-0.61	-0.45	-0.49	0.585[3]
PoNDER dataset		**[N=46]**	**[N=525]**	**[N=2808]**	**[N=4192]**	**[N=1353]**	
CORE-6D	**Mean (SD)**	0.68 (0.16)	0.78 (0.13)	0.87 (0.10)	0.91 (0.06)	0.93 (0.05)	<0.001
	ES		0.78	0.83	0.71	0.36	
SF-6D	**Mean (SD)**	0.55 (0.11)	0.63 (0.10)	0.72 (0.11)	0.79 (0.11)	0.85 (0.10)	<0.001
	ES		0.84	0.83	0.67	0.56	
Mapped EQ-5D	**Mean (SD)**	0.33 (0.29)	0.65 (0.20)	0.81 (0.15)	0.90 (0.12)	0.96 (0.09)	<0.001
	ES		1.53	1.08	0.75	0.67	
CORE-OM	**Mean (SD)**	14.41 (7.41)	10.83 (6.49)	6.80 (5.47)	3.76 (3.81)	2.39 (3.03)	<0.001
	ES		-0.55	-0.74	-0.80	-0.45	

ANOVA = analysis of variance; ES = effect size; SD = standard deviation

1. Significant for all measures in both PMS and PoNDER datasets across different levels of general health defined by question 1 of SF-12.
2. Level of significance in pairwise comparisons between different severity levels
3. Between 'excellent' and 'very good' response levels

Convergence

According to the results in Table 45, CORE-6D was strongly correlated with CORE-OM, with Pearson correlation coefficients exceeding 0.80 in all datasets, which was not an unexpected finding, given that CORE-6D is derived from CORE-OM. The generic PBMs were moderately correlated with CORE-OM, with all coefficients being within the range of 0.55-0.65. All PBMs showed

moderate to strong correlations to CIS-R, which tended to be higher for CORE-6D, with coefficients ranging from 0.53 (mapped EQ-5D) to 0.69 (CORE-6D). All correlations were significant at the 0.01 level.

Table 45. Convergence: Pearson correlation coefficients of preference-based measures with CORE-OM and CIS-R

	CORE-OM			CIS-R
	PMS dataset	**PHASE dataset**	**PoNDER dataset**	**PMS dataset**
CORE-6D	-0.83	-0.84	-0.81	-0.69
SF-6D	-0.64	NA	-0.64	-0.62
EQ-5D	NA	-0.57	NA	NA
Mapped EQ-5D	-0.55	NA	-0.61	-0.53

All correlations are significant at the 0.01 level (2-tailed); correlations are negative because CORE-OM and CIS-R clinical scores increase with increased symptom severity

Agreement between preference-based measures

As shown in Table 46, CORE-6D demonstrated rather moderate to strong agreement with generic PBMs, with ICCs ranging from 0.48 (agreement with mapped EQ-5D, PMS dataset) to 0.71 (agreement with SF-6D, PoNDER dataset). ICCs between SF-6D and mapped EQ-5D ranged from 0.73 to 0.88, indicating strong agreement that is not surprising given that both generic PBMs were derived / mapped from SF-36 or SF-12. All correlations were significant at the 0.01 level.

Table 46. Agreement: Intraclass correlation coefficients between preference-based measures

	PMS dataset		PHASE dataset	PoNDER dataset	
	CORE-6D	**SF-6D**	**CORE-6D**	**CORE-6D**	**SF-6D**
SF-6D	0.62	1	NA	0.71	1
EQ-5D	NA		0.67	NA	
Mapped EQ-5D	0.48	0.73	NA	0.68	0.88

All correlations are significant at the 0.01 level (2-tailed)

Differences in the content between CORE-6D and generic preference-based measures

Correlations between the CORE-6D and the generic PBM items were rather weak although significant in most cases (Table 47), indicating that CORE-6D may capture different aspects of HRQoL from those tapped by generic PBMs. Moderate to strong correlations (Spearman coefficients around 0.42-0.68) were found:

a. between 4 of the CORE-6D items (all except the 'risk-to-self' item and the physical item) and the 'anxiety/depression' item of EQ-5D
b. between the CORE-6D item "I feel terribly alone and isolated" and the SF-6D items on 'role limitations', 'social functioning' and 'mental health'
c. between the CORE-6D item "I am able to do most things I need to" and SF-6D items on 'role limitations' and 'social functioning'
d. between the CORE-6D physical item and
    i. all SF-6D items except the 'mental health' one;
    ii. the EQ-5D items on 'pain/discomfort' and 'mobility'

Correlations between the CORE-6D physical item and the pain items of the two generic PBMs were not as strong as it might have been expected considering the overlapping content of these items (maximum correlation coefficient observed in the datasets 0.68)

All other correlations between CORE-6D items and the items of generic PBMs were weak or non-existent. In general, all CORE-6D items correlated weakly or moderately (but in all cases significantly) with each of the 6 SF-6D items. On the other hand, CORE-6D items correlated significantly (weakly or moderately) only with some of the EQ-5D items.

Surprisingly, the CORE-6D items correlated rather weakly with the mental health item of SF-6D (except the moderate correlation between the latter and the CORE-6D item 'I feel terribly alone and isolated' - correlation coefficient 0.49), despite the focus of CORE-6D on mental health aspects of HRQoL.

Table 47. Correlations of CORE-6D items with SF-6D and EQ-5D dimensions: Spearman's rank correlations

	SF-6D dimensions: PMS dataset PoNDER dataset						EQ-5D dimensions: PHASE dataset				
CORE-6D items	**Physical functioning**	**Role limitations**	**Social functioning**	**Bodily pain**	**Mental health**	**Vitality**	**Mobility**	**Self care**	**Usual activities**	**Pain / discomfort**	**Anxiety / depression**
alone and isolated	0.17**	0.37**	0.32**	0.21**	0.38**	0.24**	0.10	-0.06	0.26**	0.13	**0.50****
	0.19**	**0.43****	**0.44****	0.22**	**0.49****	0.37**					
panic or terror	0.10*	0.39**	0.32**	0.23**	0.31**	0.20**	0.08	-0.03	0.25**	0.17*	**0.48****
	0.14**	0.30**	0.33**	0.16**	0.36**	0.25**					
felt humiliated	0.03	0.17**	0.17**	0.06	0.22**	0.09*	0.13	0.02	0.19**	0.07	**0.43****
	0.11**	0.27**	0.27**	0.14**	0.32**	0.20*					
able to do things	0.33**	**0.43****	**0.42****	0.28**	0.19**	0.33**	0.18*	0.16*	0.36**	0.11	**0.48****
	0.17**	0.30**	0.34**	0.18**	0.32**	0.32**					
plans to end life	0.10*	0.17**	0.16**	0.14**	0.14**	0.13**	0.09	0.00	0.09	-0.02	0.17*
	0.05**	0.10**	0.11**	0.06**	0.12**	0.08**					
aches, pains	**0.51****	**0.47****	**0.51****	**0.56****	0.24**	**0.48****	**0.50****	0.38**	0.33**	**0.68****	0.16*
	0.42**	0.32**	0.33**	**0.63****	0.21**	0.32**					

Notes: *=correlation is significant at the 0.05 level (2-tailed); ** = correlation is significant at the 0.01 level (2-tailed). All coefficients ≥ 0.40 are shown in bold.

Content validity

As illustrated in Table 48, the 6 items of CORE-6D capture, even if partially in some cases, all 7 major domains of HRQoL that are important to people with mental health problems, as identified by Brazier and colleagues (2014) and Connell and colleagues (2012). Some items have been assigned to more than one theme, with the secondary themes being indicated by items shown in square brackets. The theme of subjective well-being & ill-being is only partly covered by item “I feel panic or terror”, which does not cover depression and focuses on ill-being caused by panic and anxiety symptoms rather than positive well-being. More implicitly, it could be argued that the item “I feel terribly alone and isolated” indicates depressive symptoms, and can also be considered to reflect subjective ill-being. Positive aspects of activities and functioning are captured, to some extent, by the item “I am able to do most things I need to”. Close and social relationships are reflected in items “I feel terribly alone and isolated” and “I feel humiliated or shamed by other people”, respectively. The latter also reflects aspects of self-perception through the eyes of others. The item “I am able to do most things I need to” suggests autonomy and control over life. Hopelessness is captured by the item “I make plans to end my life”. Finally, the physical item of CORE-6D (“I am troubled by aches, pains, physical problems”) captures the theme of physical health.

As reported in section 2.3.3 of Chapter 2, Brazier and colleagues (2014) assessed the content validity of EQ-5D against the 7 domains of HRQoL that were identified as important to people with mental disorders. Of the 7 HRQoL domains, EQ-5D captures well the one on physical health (by items on mobility, self-care, usual activities and pain). Activity and functioning is vaguely covered by the EQ-5D item on usual activities. Social well-being, belonging and relationships is only partially captured by the EQ-5D item on usual activities. Subjective ill-being (but, like CORE-6D, not well-being) is broadly reflected in the EQ-5D item on depression and anxiety. EQ-5D is not able to represent the remaining 3 domains of HRQoL that are important to people with mental disorders, that is, control, autonomy and choice; self-perception; and hope and hopelessness.

In the same section of Chapter 2 there is also an assessment of the content validity of SF-6D against the 7 HRQoL domains that are important to populations with mental health problems, as reported by Brazier and colleagues (2014). Compared with EQ-5D, SF-6D covers a wider range of the 7 HRQoL domains that are relevant to mental health, as it is more balanced in taping both physical and mental health aspects. SF-6D captures the domain of physical health through its questions on physical functioning, vitality, and bodily pain and its sub-question on role limitations due to physical problems. Activity and functioning is covered by SF-6D items on physical functioning and role limitations. Social well-being, belonging and relationships is broadly captured by the social functioning item of SF-6D. Subjective well-being and ill-being is covered by the mental health and the vitality SF-6D items and, partly, by the sub-question on role limitations due to emotional problems. Similarly to EQ-5D, SF-6D is unable to capture the remaining 3 domains of HRQoL that are important to people with mental disorders, that is, control, autonomy and choice; self-perception; and hope and hopelessness.

In summary, the qualitative assessment of the content validity of CORE-6D, EQ-5D and SF-6D indicated that CORE-6D is more appropriate than generic PBMs in capturing aspects of HRQoL that are important to people with mental health problems.

Table 48. Content validation of CORE-6D against the main domains of health-related quality of life that are important to people with mental health problems [as identified by Brazier and colleagues (2014) and Connell and colleagues (2012)]

Domain	Sub-theme and summary description	CORE-6D item
Subjective well-being & ill-being	Distress; associated with depression, experience of psychosis and mania and anxiety Depressive mood; is associated with poor concentration, low energy and poor motivation Fear or panic and anxiety; can be caused by stressful social situations Psychosis-related: distress caused by critical voices, difficult to differentiate from reality Positive well-being: happiness and enjoyment; feeling peaceful, calm, relaxed and safe Energy and motivation (lack of both often caused by lack of sleep)	[I feel terribly alone and isolated] I feel panic or terror
Activity & functioning	Positive: work, hobbies or social interaction Negative: stressful if too demanding; fear of stress may result in avoiding enjoyable activities	[I am able to do most things I need to]
Social well-being, belonging & relationships	Relationships: close friends and family Social relationships Reactions of others – understanding, acceptance and stigma Sense of belonging	I feel terribly alone and isolated I feel humiliated or shamed by other people
Self-perception	Self-identity Self-efficacy, self-esteem and self-acceptance Self-stigma	 [I feel humiliated or shamed by other people]
Control, autonomy & choice	Dependence and independence – relating to support Self-control: mainly related to relief / management of symptoms, usually through medication Choice: money and access to resources	I am able to do most things I need to
Hope & hopelessness	Dreams and goals, involvement in activities that give meaning and purpose Hopelessness; lowering of aspirations	 I make plans to end my life
Physical health	Physical comorbidity or experience associated with mental health problem	I am troubled by aches, pains, physical problems

7.4 Discussion

This chapter evaluated the psychometric properties and content of CORE-6D, a PBM aimed at capturing HRQoL changes relating to common mental health problems, relative to those of the generic PBMs EQ-5D and SF-6D, using 3 different datasets. Where EQ-5D utility values were not directly available, they were obtained via mapping using SF-36 or SF-12 data. CORE-6D was also compared with the CORE-OM (the measure that CORE-6D was derived from) to assess the degree of the loss of information arising from moving from a 34-item instrument to a 6-item measure. Finally, CORE-6D was assessed for its content validity against the 7 areas of HRQoL that have been found to be most relevant to people with mental health problems.

Summary of findings

Comparisons between preference-based measures

CORE-6D appeared to have comparable acceptability to generic PBMs (reflected in small percentage of missing data), and showed no floor or any significant ceiling effects in populations with common mental health problems.

The responsiveness of CORE-6D was better than that of EQ-5D in the PHASE dataset, although this finding must be interpreted with caution due to small numbers of pairwise observations across different time points. In the PoNDER dataset, CORE-6D showed negligible responsiveness in the whole sample of postnatal women, but this reflected the poor performance of CORE-OM in this sample. The responsiveness of both CORE-6D and CORE-OM was much improved when data exclusively from women diagnosed with postnatal depression at baseline were analysed. Generic measures showed persistently moderate to high responsiveness in the PoNDER dataset, regardless of the sample used in the analysis.

CORE-6D demonstrated a substantially better ability to distinguish across different symptom severity groups defined by either the CORE-OM or the CIS-R score compared with generic PBMs. However, with respect to CORE-OM defined severity groups, this may be expected as CORE-6D is derived from the

CORE-OM and is made up of a subset of CORE-OM items. On the other hand, generic PBMs showed somewhat better (and, in contrast to CORE-6D, always significant) ability to distinguish across groups with different levels of general health as determined by responses to question 1 of the SF-12.

CORE-6D utility values showed better correlation with CORE-OM than any other generic PBM did, but this was anticipated, considering its derivation from CORE-OM and its sharing of common items. On the other hand, the correlation of all PBMs with CIS-R, which was used as an indicator of convergent validity, was comparable. Agreement between CORE-6D and the generic PBMs was moderate to strong. Regarding the correlations between individual items of CORE-6D and items of generic PBMs, all CORE-6D items correlated weakly to moderately with each of the SF-6D items; on the other hand, the only correlations between CORE-6D and EQ-5D items were moderate and were observed between the CORE-6D emotional items and the depression/ anxiety item of EQ-5D, and also between the CORE-6D physical item and the EQ-5D items on mobility and pain/discomfort.

In contrast to generic PBMs, CORE-6D appeared to broadly cover all 7 major themes of HRQoL that have been found to be important to people with mental disorders, although some important sub-themes, such as subjective well-being, were not captured by any of the CORE-6D items.

Extent of loss of information relative to the CORE-OM

As anticipated, CORE-6D correlated strongly with the CORE-OM; nevertheless, there was a modest reduction in the responsiveness of CORE-6D relative to that of the CORE-OM, suggesting some loss of sensitivity. Perhaps not surprisingly, CORE-OM showed a substantially higher ability than CORE-6D in distinguishing across different symptom severity groups when symptom severity was determined using the CORE-OM clinical score. However, the two measures were more comparable in distinguishing across different symptom severity groups determined by CIS-R. The ability of the two measures in distinguishing across different levels of general health was similar, with CORE-6D showing better ability at the lower levels of health, while both measures seemed unable to capture small differences between adjacent high

levels of general health (as measured by responses to question 1 of the SF-12). These results overall suggest a rather small loss of information at the derivation of CORE-6D from CORE-OM.

Interpretation of the results

The results indicate that the performance of CORE-6D overall compares to that of generic PBMs. CORE-6D showed, as expected, higher ability to distinguish across groups with different levels of mental symptom severity, and lower ability to distinguish across groups with different levels of overall health. It also showed higher responsiveness that EQ-5D in the PHASE dataset. The low responsiveness of CORE-6D and CORE-OM that was observed in the PoNDER dataset can be possibly explained by the fact that the majority of women in the sample (about 83%) did not have postnatal depression at baseline and therefore had little scope for improvement in mental health symptoms which are the focus of CORE-6D and CORE-OM. On the other hand, the moderate responsiveness of the generic PBMs in the whole sample can be justified on the basis of evidence that childbirth is associated with numerous physical problems, accompanied by extreme tiredness and exhaustion due to the demands of parenting, which naturally resolve over time (Bick & MacArthur, 1995; Brown & Lumley, 1998; Thompson et al., 2002); such problems most likely reduce the baseline HRQoL in terms of physical and social functioning, usual activities, self-care and vitality, i.e. dimensions that are the main focus of generic PBMs. It therefore appears that SF-6D and mapped-EQ-5D apparently captured the scope for natural improvement in the physical health of all postnatal women over the time period following childbirth. The increased responsiveness of all measures (in particular of the condition-specific ones) in the sub-sample of women with postnatal depression may suggest that the condition-specific measures captured the improvement in mental health symptoms of the depressed women following treatment, whereas the changes in the generic PBMs over time reflected the improvements in overall health experienced by the women in this sub-group.

The lack of strong correlations between individual items of CORE-6D and items of generic PBMs suggests that CORE-6D may capture different aspects

of HRQoL from those reflected in items of generic PBMs, even though the content of some CORE-6D items appears to overlap with the content of some of the items of the generic PBMs. The lack of strong correlations between seemingly similar items of CORE-6D and generic PBMs may be attributable to

i. differences in the context, as items are answered within a different set of questions, with CORE-6D items being embedded in a large, mental health-specific questionnaire and SF-6D items being embedded in a 12-item or a 36-item generic HRQoL instrument (while EQ-5D forms a distinct questionnaire)
ii. differences in the wording of the items
iii. differences in the recall period related to each measure: CORE-6D measures health over the past week, SF-6D asks questions mostly over the past 4 weeks, whereas EQ-5D measures health today.

Nevertheless, the most important differences between CORE-6D and generic PBMs lie in the actual content of their items, as revealed by qualitative assessment of their content against the 7 themes identified as relevant to people with mental health problems, with CORE-6D covering fully or partially all 7 areas, while generic PBMs were able to represent only 4; this is in line with the conclusion of Brazier and colleagues (2014), who noted that the factors that determine the HRQoL in people with mental health problems are quite complex and generic measures are not able to capture the aspects of HRQoL that matter most to this population due to their focus on physical aspects of health.

Limitations of the analyses

Analyses described in this chapter were conducted on 3 separate datasets. This increased the number of analyses that could be performed and allowed checking of the reproducibility of results in different populations. The selection of these 3 datasets was dictated by the need for concurrent availability of CORE-OM and generic PBM data in the same dataset. However, of the 3 datasets, only the PHASE dataset included a ‘clinical’ population seeking treatment for common mental health problems. The PMS dataset included participants in a survey, a considerable proportion of whom had no indication

for or symptoms of a mental health problem; in the PoNDER dataset only a minority of women were diagnosed with postnatal depression at baseline. Consequently, with the exception of the PHASE dataset, the dataset populations are not fully representative of people with common mental health problems, which is the target patient population of the CORE-6D. A further limitation was that, with the exception of CIS-R in the PMS dataset, the different severity sub-groups used to test the known groups validity of PBMs were determined by CORE-OM clinical scores, due to unavailability of alternative mental health-specific instruments in the datasets. This is likely to have introduced bias in the analyses, as CORE-6D is directly related to the CORE-OM; therefore its superior performance in terms of known groups validity across symptom severity groups relative to generic PBMs is not surprising. On the other hand, it should be noted that CORE-6D demonstrated a higher known groups validity relative to EQ-5D and SF-6D even when CIS-R was used to form different symptom severity groups.

Another limitation of the analyses was that responsiveness to change over time was assumed to reflect underlying important changes in health and HRQoL following treatment. However, it is not known whether treatment was effective in the study samples and, if it was, to what extent, and therefore responsiveness has been measured on the assumption and expectation (rather than real observation) that people's health and HRQoL did change overtime following treatment. This assumption was necessary due to lack of a gold standard measure that could confirm underlying differences in health and HRQoL overtime.

Finally, validation of CORE-6D was attempted by comparing its performance to that of generic PBMs. However, since the use of generic PBMs in mental health population is, as already discussed in Chapter 2, problematic, comparison of CORE-6D with generic PBMs is not ideal. On the other hand, quality of life is a subjective as well as complex concept, for the measurement of which no acceptable gold standard is available at the moment (Brazier et al., 2014). Since a more relevant and reliable comparison was not possible, assessment of the properties of CORE-6D relative to those of generic PBMs

was considered acceptable. It is acknowledged, though, that, although such comparisons may provide some indications on the psychometric performance of CORE-6D, they cannot lead to unambiguous results that prove or discard the measure's value as a mental health-specific PBM.

7.5 Conclusion

The analyses presented in this chapter show promising results regarding the acceptability, validity and responsiveness of CORE-6D as a condition-specific PBM, although no firm conclusions on the psychometric properties of CORE-6D can be drawn at this stage. CORE-6D shows acceptable agreement with generic PBMs and suffers from relatively small loss of information compared with the CORE-OM. Its items can broadly capture all HRQoL aspects that are important to people with mental health problems, although it is acknowledged that each aspect covers a broad area of HRQoL and includes several sub-themes, the majority of which are not reflected in CORE-6D items. However, any limitations regarding loss of information and inability to capture a number of HRQoL sub-themes that are relevant to people with mental disorders might be deemed to be acceptable, considering that CORE-6D is by intension a brief, concise 6-item measure. The findings on the psychometric properties of CORE-6D are constrained by a number of limitations. Nevertheless, the properties of CORE-6D suggest that it may be appropriate to use as a utility measure in economic evaluations of interventions for common mental health problems. Further research should explore the psychometric properties of CORE-6D in more depth and establish its value as a PBM for people with common mental health problems.

Chapter 8. Discussion and conclusion

8.1 Introduction

The subject of this thesis was the development of CORE-6D, a PBM that is relevant to common mental health problems, derived from a large CSM, the CORE-OM. Due to the high correlation between the domains and items of CORE-OM, a novel methodology was developed and applied for the derivation of CORE-6D, rather than standard methods described in the literature previous to this research, which have been employed for the development of typically multidimensional PBMs.

The methodology adopted in the development of CORE-6D as well as its strengths and limitations have been described in previous chapters. The purpose of this final chapter is to evaluate the contribution of this thesis to the broader methodology described in the literature for the derivation of PBMs from existing measures, discuss the implications for policy resulting from the output of this thesis, point to some broader issues on the role of condition-specific PBMs in the wider healthcare resource allocation environment and the potential role of patients' preferences in the economic assessment of healthcare interventions, and propose areas for further research.

8.2 Contribution of this thesis to the methodology for deriving preference-based measures from existing instruments

The standard methodology that is described in the literature for the derivation of health state classifications from existing, longer measures, the subsequent selection of health states for valuation and the modelling techniques used to attach appropriate utility values to all health states described by the health state classification presupposes independence between the dimensions of the original measure. In summary, the standard methodology for the derivation of health state classifications comprises identification and/or establishment of the dimensions of the original measure using FA or PCA, followed by selection of the most suitable item(s) of each dimension based on psychometric criteria, judgement and, more recently, Rasch analysis. This process typically results in

the development of a health state classification where each item represents one dimension and items are not considerably correlated to each other. For the selection of the health states to be valued and for the modelling of valuation data in order to produce an algorithm linking all possible health states to a utility value, the two main approaches proposed in the literature are the composite, which uses statistical designs for the selection of a sample of health states for valuation, followed by regression modelling techniques for the prediction of utility values for all health states described by the health state descriptive system; and the decomposed, based on MAUT, which involves valuation of individual dimensions and of a number of multidimensional states followed by the development of a system of equations that allow attaching utility values to every multidimensional health state. Both approaches require dimensions to be independent and not highly correlated at the descriptive level.

The novel methods proposed in this thesis for the derivation of new PBMs from existing instruments are appropriate to apply to unidimensional measures or measures with highly correlated dimensions, in order to overcome problems that would arise from employment of conventional techniques: one such problem could be the generation of implausible health states when using statistical designs; another implication from the use of standard techniques could be the need for complex econometric modelling to take account of multiple interaction effects between highly correlated items when modelling utility values. The proposed methodology, which avoids this kind of problems, is likely most applicable to CSMs that have a narrow scope, for example by focusing mainly on symptoms or one aspect of patients' HRQoL. In all cases, the dimensionality of the existing measure should be examined at an initial stage of the process; exploratory FA can be used for this purpose, to give an indication of the extent of unidimensionality and the number of dimensions covered by the original instrument (Tennant & Pallant, 2006).

If the original instrument is clearly multidimensional, then standard methodology is appropriate to use, as described in the relevant literature (for example, Brazier et al., 2002 & 2008; Young et al., 2009 & 2011). If, on the

other hand, the original instrument is found to have a largely unidimensional component or highly correlated dimensions, then Rasch analysis can be used to select items in order to construct a unidimensional health state classification (or a health state classification with a large unidimensional component) and to select plausible health states for valuation by inspection of the Rasch item threshold map. Subsequently, if the new health state classification comprises a unidimensional scale, then the approach described by Young and colleagues (2010) can be adopted in order to predict utility values for all potential health states using the results of Rasch analysis. If, on the other hand, the new health state classification has more dimensions but with a prevailing unidimensional component, then the hybrid approach developed for this thesis can be used for modelling utility values following the valuation of plausible health states.

The methodology developed for this thesis regarding the derivation of health state classifications from instruments without clear multidimensional structure has already been adopted by a number of studies that were undertaken after the work conducted for this thesis was reported. The earliest of these studies was carried out by Young and colleagues (2010), who used Rasch analysis in order to derive a health state descriptive system from the Flushing Symptoms Questionnaire (FSQ), a unidimensional scale measuring symptoms associated with flushing as a side effect of taking niacin medications. Following construction of a unidimensional health state classification based primarily on Rasch analysis, the authors identified plausible health states by inspection of the Rasch item threshold map and undertook a valuation survey. Up to this point, the study adopted the approach proposed in this thesis. The authors developed this approach further, as their study was the first to examine the relationship between the utility values obtained from the valuation survey and the respective Rasch logit values of the health states valued, and to use this relationship in order to model utility values for all health states described by the PBM.

Kowalski and colleagues (2012) adopted the methodology developed for this thesis to derive a unidimensional health state classification from the National Eye Institute Visual-Function Questionnaire-25 (NEI VFQ-25). A number of

health states were subsequently included in a valuation survey, followed by regression modelling in order to attach utility values to all health states described by the classification (Rentz et al., 2014). The method for selection of health states for the valuation survey was not reported, but it was likely based on the methodology proposed in this thesis, as subsequent modelling of utility data followed the method developed by Young and colleagues (2010). The authors justified adopting this modelling approach "*because the dimensions of the health state descriptions* [were] *not independent*" and therefore "*conventional methods* [for modelling valuation data] *could not be used*".

Versteegh and colleagues (2012) also used Rasch analysis as proposed in this thesis to derive a unidimensional health state classification from the Health Assessment Questionnaire (HAQ), a measure widely used in rheumatology to assess functional abilities. Selection of health states for valuation was made using an orthogonal block design combined with a selection of the most observed health states and modelling of valuation data followed standard statistical methodology, without inclusion of any interaction terms. The authors reported that "*the unidimensionality of the HAQ caused some problems in the valuation task*" because items of the classification were highly correlated resulting in one of the health states selected for valuation being implausible and causing "*confusion with some of the respondents*". Use of the Rasch item threshold map for the selection of health states would have prevented this situation, as it would have led to the identification of plausible health states for inclusion in the valuation survey.

In the same publication (Versteegh et al, 2012), the authors reported that they derived a 2-dimensional health state classification from the Multiple Sclerosis Impact Scale 29 (MSIS-29), an instrument assessing the physical and psychological impact of multiple sclerosis. MSIS-29 consists of a physical and a psychological scale and Rasch analysis was undertaken separately on each scale, to create a 2-dimensional health state classification consisting of two unidimensional components. The authors employed standard techniques for the selection of health states (combined with a selection of the most observied health states) and for modelling utility values for all states described by the

health state classification, without reporting any problems regarding the plausibility of the health states selected for the valuation survey or the need for complex modelling to account for interaction between items within each of the 2 unidimensional components of the new measure.

In addition to those studies, Sundaram and colleagues (2009) developed in parallel similar methodology with that proposed in this thesis in order to derive a health state classification from the Audit of Diabetes-Dependent Quality-of-Life (ADDQoL), a 18-item instrument measuring HRQoL in patients with diabetes. As part of the process, Rasch analysis was undertaken on the whole measure aiming to develop a unidimensional instrument amenable to valuation. The resulting health state classification, the Diabetes Utility Index (DUI) was reported to be unidimensional and, at the same time, to consist of 5 distinct attributes (physical ability and energy level, relationships, mood and feelings, enjoyment of diet and satisfaction with managing diabetes). Valuation of DUI was achieved following application of MAUT using a multiplicative model that took account of the preference interactions between the attributes (Sundaram et al., 2010).

Overall, the literature suggests that the methodology developed for this thesis can be useful (and has already been used) in the development of PBMs that are derived from instruments with high correlations between their items, both in the construction of the health state classification and in the selection of plausible health states for valuation.

8.3 A new preference-based measure for cost-utility analysis of mental health interventions – implications for mental health policy and practice

The output of this thesis, CORE-6D, is a 2-dimensional PBM, consisting of a 5-item emotional component and a physical item. It has been developed following Rasch analysis and psychometric testing on CORE-OM data from people with common mental health problems presenting to NHS primary care services, and has been validated on a large mixed sample of people presenting to either primary or secondary services in the UK. Therefore, it is suitable for use in a wide range of services and settings, and can capture the

full spectrum and range of symptom severity of common mental health problems, as ensured by the Rasch analysis criteria used at its development.

8.3.1 Comparison of the psychometric properties of CORE-6D with those of generic preference-based measures

By design, CORE-6D appears to be more suitable than generic PBMs such as EQ-5D, SF-6D and HUI-3 for the estimation of QALYs in cost-utility analyses undertaken in the area of mental health. With 5 out of its 6 items representing emotional aspects of HRQoL, CORE-6D appears to be more relevant to people with mental disorders, compared with EQ-5D, which consists of 4 items on physical health (mobility, self-care, usual activities and pain/discomfort) and one mental health item (anxiety/depression). Similarly, HUI-3 contains 6 'physical health' attributes (vision, hearing, speech, ambulation, dexterity and pain), one attribute on cognition, and only one on emotion. SF-6D on the other hand, although generic, is somewhat more balanced between physical and emotional aspects of HRQoL, with 3 exclusively 'physical health' dimensions (physical functioning, bodily pain and vitality), one pure 'mental health' dimension, and 2 dimensions relating to both physical and mental health (role limitations and social functioning).

Indeed, as discussed in Chapter 2, the report by Brazier and colleagues (2014), which attempted to validate generic PBMs in a range of mental disorders, described a mixed picture of the performance of EQ-5D and SF-6D regarding their psychometric properties, which depended also on the type of mental disorder assessed. Overall, EQ-5D and SF-6D showed good construct validity and responsiveness in the area of common mental health problems, which is the focus of CORE-6D. Generic PBMs appeared to perform satisfactorily in depression, but less so in anxiety and personality disorders. Results were mixed in schizophrenia and bipolar disorder. The authors suggested that generic PBMs may be picking depressive symptoms (or comorbid depression) rather than core symptoms associated with a range of conditions, including anxiety and schizophrenia. In dementia, self-reported EQ-5D had questionable validity and poor agreement with proxy EQ-5D ratings; the later appeared to have higher validity (Hounsome et al., 2011).

The review of qualitative evidence on aspects of HRQoL that are important to people with mental health problems suggested that generic PBMs fail to address the complexity of quality of life measurement and the broad range of domains that are important to people with mental health problems. Based on their findings, the authors concluded that none of the existing generic measures can adequately capture the aspects of HRQoL that are important to people with mental health problems and proposed the development of a new mental health-specific PBM that is relevant across all populations with mental health problems. The authors acknowledged that it may not be possible to capture all dimensions of physical and mental health with the same detail in the new measure, but the latter would need to incorporate the impact of both physical and mental health problems.

As reported in Chapter 7, psychometric analyses of CORE-6D data showed promising results regarding the measure's responsiveness and construct validity in populations with common mental health problems, although these analyses had a number of limitations and further research needs to validate the findings. In contrast to generic PBMs, CORE-6D is able to broadly tap the 7 major themes of HRQoL that were found to be most relevant to people with mental disorders, namely, subjective well-being & ill-being; activity & functioning; social well-being, belonging & relationships; self-perception; control, autonomy & choice; hope & hopelessness; and physical health (Brazier et al., 2014; Connell et al., 2012). It is true that CORE-6D focuses largely on emotional symptoms, captured by its 5 emotional items, but it also incorporates the impact of physical problems on HRQoL, as it includes a physical item. The composition of CORE-6D reflects the structure of the original measure CORE-OM, which has been designed primarily for the monitoring of emotional, rather than physical, symptoms. Inclusion of one physical item in CORE-6D allows a rather crude representation of physical symptoms, which, nevertheless, enables the assessment and valuation of both emotional and physical dimensions of HRQoL in people with common mental health problems.

8.3.2 Use of CORE-6D in cost-utility analysis of mental health interventions – advantages, limitations and implications

Cost-utility analysis has proved to be problematic in the area of mental health. Generic PBMs that would allow estimation of QALYs are not routinely used in mental health clinical practice or in the design of clinical and/or economic studies, possibly revealing unacceptability of such measures among mental health practitioners, patients and researchers (Crawford et al., 2010; Gilbody et al., 2003). A review of relatively recent NICE guidelines in the area of mental health, which included systematic reviews of economic evaluations of pharmacological and psychosocial mental health interventions, suggests that less than 50% of the economic evaluations in this area are in the form of cost-utility analysis. More specifically, in the area of depression, only 13 out of the 29 economic evaluations published between 1998-2008 that were included in the respective NICE guideline were cost-utility analyses and only 9 of them reported use of a generic PBM (National Collaborating Centre for Mental Health, 2010a). In the area of GAD, 2 of the 5 economic analyses published between 1997-2009 that were included in the NICE guideline were cost-utility analyses (National Collaborating Centre for Mental Health, 2011), while in the area of social anxiety only one out of the 4 economic evaluations that were included in the NICE guideline used the QALY as the measure of outcome (National Collaborating Centre for Mental Health, 2013). In more severe mental disorders such as bipolar disorder and schizophrenia the proportion of cost-utility analyses in economic evaluations considered alongside NICE guideline development was 6/14 (National Collaborating Centre for Mental Health, 2014) and 10/35 (National Collaborating Centre for Mental Health, 2010b), respectively. In particular in the area of schizophrenia only 3 out of the 10 cost-utility analyses utilised a generic PBM, with the remaining 7 using utility values estimated based on vignettes or a CSM. This latter finding may reflect the perceived unsuitability of using generic PBMs in the area of schizophrenia. It should be noted that a number of the NICE guidelines reviewed above were published a few years ago, and the proportion of economic studies in the form of cost-utility analysis as well as the use of generic PBMS for estimation of

QALYs in the areas covered by the guidelines is likely to have increased in more recent years.

The majority of economic studies identified in the review of NICE clinical guidelines were cost effectiveness or cost consequence analyses, where the outcome measure was often expressed as ‘proportion of people responding to treatment’, ‘number of relapses avoided’, ‘number of depression-free days’, and so on. Use of such measures in economic evaluation may reflect unacceptability or perceived inappropriateness of generic PBMs in the area of mental health, or even lack of a more appropriate condition-specific PBM. However, use of natural units as outcome measures in economic studies limits comparability across disease areas (even within mental health) and requires outcome-specific judgements when determining cost effectiveness. A more characteristic example is when the measure of outcome is expressed as a change score on a continuous scale, so that the ICER is estimated as cost per change in score, making judgements on cost effectiveness dubious or even impossible (for example McCrone et al., 2009, who evaluated the cost effectiveness of computerised self-help in people with agoraphobia/panic disorder). The limitations arising from use of outcome measures other than the QALY highlight the need for use of a valid, responsive and acceptable PBM that can be used for estimation of QALYs in the area of mental health.

In this context and given the lack of contect validity of generic PBMs in mental health, CORE-6D can be used to conduct cost-utility analyses for the assessment of interventions and programmes for common mental health problems. Results of such analyses will be more readily interpretable compared with economic analyses that use a natural unit as the measure of outcome; moreover, they allow comparisons with results of cost-utility analyses in other areas of mental health. The value of CORE-6D as a PBM is increased considering the wide use of CORE-OM (and, consequently, CORE-6D) for the clinical monitoring of people with common mental health problems in the UK practice setting. CORE-OM is an acceptable measure to both patients and healthcare professionals; moreover it is freely available to users (Barkham et al., 2001). Derivation of utility values from CORE-6D does not place any

burden on patients in terms of answering extra questionnaires when CORE-OM data are being recorded. Relevant syntax can be read by SPSS data files to estimate utilities directly, when CORE-OM data are available, so there is no extra burden to researchers or clinicians either.

One limitation of CORE-6D in the wider mental health services context is that it is probably not suitable to use as a universal PBM across all mental health conditions. This is because the original instrument, CORE-OM, has been designed for the measurement of psychological distress primarily associated with common mental health problems; these include various forms of unipolar depression and anxiety disorders such as GAD, panic disorder, phobias, and OCD. CORE-OM covers aspects of severe mental illness, as it includes items capturing suicidal thinking, self-harm, threatening behaviour and other severe distress. However, CORE-OM was not specifically designed for people with a psychotic element, and therefore may not be valid for outcome measurement in people with psychosis, including people with schizophrenia, bipolar disorder and personality disorders. Consequently, CORE-6D may not be appropriate for the estimation of QALYs in the evaluation of interventions targeted at psychotic disorders and therefore cannot be used as a generic mental health PBM. Nevertheless, common mental health problems constitute the most prevalent group of mental disorders, experienced by 17.6% of people aged 16 to 64 years in England; for comparison, the prevalence of psychotic disorders in this population is 0.4% (McManus et al., 2009). It can be thus concluded that CORE-6D is suitable for the estimation of QALYs in the large majority of people with mental disorders, including those with common mental health problems and, potentially, those with more severe mental illness, such as, for example, people self-harming. It is acknowledged that the purpose of this thesis was to develop a generic mental health PBM that can be used across all mental health areas. However, currently, an outcome measure that is valid and responsive in capturing symptoms and aspects of HRQoL that are relevant across all mental disorders, which could be used to derive a generic mental health PBM, does not seem to exist. Therefore, a generic mental health PBM would need to be developed *de novo*, and this is proposed in section 8.6.7.

One point worth noting is that CORE-6D is not an autonomous measure; its 6 items are embedded in the 34-item CORE-OM. Reading the 6 items of CORE-6D within the context of the CORE-OM questionnaire may effect the meaning, intensity and relative importance of each of the 6 items, which, in turn, may have an impact on the responses obtained. Although there is evidence that the response rates and quality of responses to instruments embedded in longer questionnaires are not affected by the length of the questionnaire (Jenkinson et al., 2003), it is not known whether and how the context of the longer questionnaire affects responders, and whether isolation of a small number of items (such as the items of CORE-6D) out of the context of a longer questionnaire (such as the CORE-OM), has any effect on rates and levels of responses. Use of CORE-6D as an independent measure in a study that does not use CORE-OM may be appealing, given the measure's brevity and function as a utility measure, but may in theory elicit different responses from those obtained when all 34 items of CORE-OM are included in the questionnaire.

Indeed, a study that compared SF-6D utility values generated from responses to the SF-36 with utility values obtained from the SF-6D administered as an independent instrument demonstrated that there were significant differences between the two sets of values (Ferreira et al., 2013). The authors concluded that since the SF-6D was originally designed to derive utilities from the SF-36, it should be used in this context and not as an independent measure. Further research should address the same issue for CORE-6D, as proposed in section 8.6.5. It should be noted, though, that the SF-6D utility values generated from the SF-36 are by design determined by responses to 11 SF-36 items, as some SF-6D items are scored by combining responses to 2-3 SF-36 items (Brazier et al., 2002), whereas SF-6D utility values obtained from the 'independent' SF-6D were generated using direct responses to the 6 SF-6D items (Ferreira et al., 2013). This difference in the number/content of items expressed in the two SF-6D versions might be responsible, at least to some degree, for the discrepancy in the 2 sets of utility values obtained. In contrast, the scoring of the 6 CORE-6D items depends exclusively on the scoring of these particular items within the CORE-OM, so, apart from differences in response levels and

small changes in wording, the 6 items of an 'independent' CORE-6D are directly comparable to the respective 6 items within the CORE-OM; therefore, the possibility that CORE-6D can stand as an independent PBM is likely higher compared with that of SF-6D.

Given the routine use of CORE-OM in the clinical monitoring of people with common mental health problems in the UK, CORE-6D is expected to contribute to the wider assessment of healthcare interventions for the management of common mental health problems in the form of cost-utility analysis in the UK, using existing and prospective CORE-OM datasets, in particular those that include the CORE-OM but no generic PBMs. Beyond the UK, CORE-OM can be routinely used in many other countries, as validated translations are now available in about 20 other languages, including Welsh, Norwegian, Spanish, Portuguese, German, Dutch, Greek, Italian, Danish, Icelandic, Swedish, Polish, Finnish, Lithuanian, Slovak, Turkish, Croatian, Albanian and Gujarati[4]. Use of CORE-6D as a PBM in settings that use a translated version of CORE-OM is a possibility, although further work is needed, as discussed in section 8.6.6.

The usefulness and anticipated contribution of CORE-6D in the economic assessment of mental health interventions and programmes is evident. However, when decisions accruing from economic evaluations in the area of mental health have a knock-on effect on other areas of healthcare, for example in the wider health policy-making context such as that of NICE, the appropriateness of using CORE-6D may be questioned, due to expressed concerns about the limitations of condition-specific PBMs and doubts regarding their comparability with generic PBMs. These issues are discussed in the next section.

[4] http://www.coreims.co.uk/About_Core_Translations.html [Accessed 25 April 2013].

8.4 Comparison of condition-specific preference-based measures with generic ones

Is there a place for CORE-6D in the current UK healthcare decision-making context?

The usefulness of CORE-6D, as well as of any other condition-specific PBM, in the wider NICE decision-making context is more controversial. As discussed in Chapter 1, the necessity for the development of condition-specific PBMs derives from the inappropriateness or insensitivity of generic PBMs in capturing relevant HRQoL aspects in various medical conditions and patient populations – and this was the rationale for the development of CORE-6D as well. In addition to being more relevant and sensitive, condition-specific PBMs are more likely to be acceptable to patients and clinicians, and, if derived from an existing measure that is being routinely used, they do not require extra time for their completion. Condition-specific PBMs can be used in prospective but also retrospective analyses when data on generic measures are lacking (Brazier et al., 2007).

8.4.1 Limitations of condition-specific preference-based measures

Apart from their apparent advantages, condition-specific PBMs are characterised by a number of limitations, which may considerably reduce their comparability with generic PBMs. An important flaw of the condition-specific PBMs is that they normally have a narrow scope and thus capture a limited number of HRQoL dimensions. For example, they normally do not consider side-effects of treatment or comorbidities. Omission of comorbidities from the health state descriptive system is likely to distort the results of a valuation survey when there is preference interaction, that is, when the impact of comorbidities (or, indeed, of any other dimensions) on preferences is not simply additive (Brazier & Tsuchiya, 2010; Rowen & Brazier, 2011). It is true that some degree of interaction has been shown to exist for generic PBMs (Brazier et al., 2002; Dolan, 1997; Feeny et al., 2002). But for condition-specific PBMs, the impact of such an interaction may be even more significant, given their narrower scope, which entails omission of a wider range of dimensions from their descriptive system.

Indeed, there is evidence that addition of extra dimensions to condition-specific PBMs has had a significant impact on mean utility values, although not always in the expected direction; the impact of adding an extra dimension may not be additive and may not be consistent across existing dimensions, resulting to the need for re-valuation when addition of an extra dimension to a condition-specific CSM is decided in order to capture side effects or comorbidities (Brazier et al., 2011 & 2012). Inclusion of extra dimensions on PBMs derived from CSMs limits the value of deriving the PBM from an existing, readily available measure, as it requires collection of additional data relating to the extra dimension. Furthermore, extra dimensions may miss unknown side effects or less frequent comorbidities. For this reason, researchers have proposed the addition of extra dimensions to generic PBMs to make them more relevant to specific conditions, as an alternative to use of condition-specific PBMs (Rowen & Brazier, 2011). Examples of such add-ons include the addition of a sleep dimension to EQ-5D (Yang et al., 2013a) as well as the addition of a cognitive dimension to the same measure (Krabbe et al., 1999). However, this solution has also limitations as it requires inclusion of the extra questions in prospective studies, and does not allow estimation of utilities (and QALYs) from existing datasets. Furthermore, there is only scope for bolting-in one or two extra dimensions before the instrument becomes too large for valuation (Brazier et al., 2007).

Another flaw of condition-specific PBMs relative to generic ones relates to the distortions created in the valuation process by focusing effects, i.e. when respondents overrate the importance of the symptoms associated with the condition being described because they are provided with a narrow perspective of HRQoL. Such effects may potentially generate a larger decrement than a generic PBM because the respondent is not being given the broader HRQoL context, and may have important implications for omitted comorbidities (Brazier & Tsuchiya, 2010). However, it is possible that respondents do implicitly consider other dimensions of HRQoL when valuing narrow health states described by condition-specific PBMs, and therefore the

impact of focusing effects may be less substantial than initially thought (Brazier et al., 2012).

Similar to focusing effects is the impact of naming the condition, which is usually inherent in condition-specific PBMs; explicitly stating the condition that is being valued may potentially distort respondents' preferences on HRQoL due to their preconceptions about the condition or their experience of the condition as patients or carers (Brazier & Tsuchiya, 2010; Brazier et al., 2012). Here, research findings are mixed, as the effect on utility values appears to depend on the label used and the severity of the health state being valued (Brazier et al., 2012). One way to limit this effect would be to remove the label from the PBM, but in the case of PBMs derived from existing CSMs this would result in discrepancies between the PBM and the original CSM. Alternatively, a *de novo* condition-specific PBM that does not name the condition could be developed. A final solution would be to retain the label and accept distortions in valuation, a solution that may actually lead to higher accuracy in the health state description and thus in the resulting utility values, although this issue warrants further research (Brazier et al., 2012).

A final issue relating to the comparability across PBMs (either generic or condition-specific) is the upper anchor used in valuation. If this upper anchor is the instrument-specific best state (which was the case in the valuation of generic PBMs), then respondents may not necessarily assume that other, omitted dimensions are at their optimum level and instead they may imagine other health problems, perhaps their own health, which could potentially affect their preferences. In order to ensure comparability across PBMs, it has been recommended that a generic 'full health' upper anchor be used across valuations of PBMs (Brazier et al., 2012).

8.4.2 NICE position on the use of condition-specific preference-based measures

Comparability across different PBMs is crucial when economic evaluations using a variety of PBMs are undertaken to inform decisions within the same resource allocation context. To enhance comparability and consistency across its appraisal programme, NICE has explicitly expressed a preference for the

EQ-5D for use in cost-utility analyses of interventions for adults (National Institute for Health and Care Excellence, 2013). Nonetheless, NICE recognises that EQ-5D data may not be available or may be inappropriate for the condition or effects of treatment. When EQ-5D data are not available, NICE recommends mapping of other available measures on EQ-5D. In situations where the use of EQ-5D is considered to be inappropriate, NICE requires reporting of quantitative and qualitative empirical evidence on the lack of responsiveness, construct and content validity of the EQ-5D in the particular patient population. Where mapping onto EQ-5D is not possible or the use of EQ-5D is inappropriate, alternative PBMs may be used, accompanied by a detailed reporting of the methods used for their development, evidence of their validity, and description of the impact of methods on the resulting utility values. Moreover, NICE requests information on the extent of the impact of the use of the alternative measure on the value of the QALYs gained. For any PBM used to inform the Institute's decisions, NICE requires that the measurement of changes in HRQoL be reported directly from patients (or, if this is not possible, by persons acting as their carers in preference to healthcare professionals), and the respective utility values be based on public preferences, elicited from a representative sample of the UK general population using a choice-based method (i.e. TTO or SG) (National Institute for Health and Care Excellence, 2013). These rules for the valuation of newly developed PBMs aim to ensure a better degree of comparability with EQ-5D, and, thus, consistency across the Institute's appraisal programme.

8.4.3 The place of CORE-6D in the NICE decision-making context

Existing psychometric evidence indicates that EQ-5D may be appropriate to use in some mental health conditions, mainly depression and, to some extent, anxiety and personality disorders, but its use is problematic in more severe mental disorders such as schizophrenia and bipolar disorder. The performance of EQ-5D in some common mental health problems such as OCD, panic disorder, generalised and specific phobias has not been assessed due to lack of relevant evidence. Qualitative evidence suggests that EQ-5D lacks important content validity in people with mental disorders (Brazier et al., 2014). EQ-5D is quite often absent from clinical evaluations carried out in mental

health. Attempted mapping of a number of widely used CSMs (including routinely used measures of common mental health problems such as the Hospital Anxiety and Depression Scale [HADS], PHQ-9, GAD-7, GHQ-12 and CORE-OM) onto generic measures (mainly SF-6D but also EQ-5D in lesser extent) using datasets of populations with depression and/or anxiety revealed weaknesses in the mapping functions (Brazier et al., 2014), which may indicate that generic PBMs are unable to fully capture the range of symptoms experienced by people with mental health problems. Overall, this evidence suggests that EQ-5D may not be appropriate to use in some populations with mental health problems, and points to the direction of the development and use of an alternative PBM in such populations.

The analyses presented in Chapter 7 showed promising results regarding the validity and responsiveness of CORE-6D in populations with depression and/or anxiety, although further research needs to validate these findings and expand analyses to a wider range of populations with common mental disorders, which are the focus of CORE-6D; moreover, qualitative analysis indicated that CORE-6D has higher content validity compared with generic PBMs in populations with common mental health problems. The methods used for the valuation of CORE-6D are comparable to those used at the development of the EQ-5D utility index, i.e. utility values were elicited from a random sample of the UK population using the MVH group TTO protocol (Dolan et al., 1996; MVH Group, 1995). Consideration of the issues relating to the performance and availability of EQ-5D in populations with mental health problems and the promising properties of CORE-6D appear to support the use of CORE-6D in the NICE policy context for the evaluation of interventions and programmes for people with common mental health problems, in situations where EQ-5D is shown not to perform satisfactorily or is not available. The results presented in Chapter 7 also provide some information on the potential impact of the use of CORE-6D instead of EQ-5D regarding the expected mean change in utility values and standard deviation over time and across different symptom severity groups, which is required by NICE before a new PBM can be used in a cost-utility analysis conducted as part of the Institute’s appraisal programme. According to these results, CORE-6D showed smaller mean changes in utility

values over time and across different severity groups compared with EQ-5D and SF-6D, but overall larger values of SRM and ES (due to smaller standard deviation), which is important for the power of a study as it indicates that CORE-6D may be able to detect significant differences with a smaller sample size. Moreover, this may lead to reduced uncertainty and affect the output of probabilistic sensitivity analysis in economic evaluation (Brazier et al., 2012). It needs to be emphasised, though, that before CORE-6D is used in this context, it is advisable that its psychometric properties are validated in larger datasets and other populations with common mental health problems.

As a condition-specific PBM, CORE-6D suffers from the limitations described in section 8.4.1, including omission of side effects and comorbidities, focusing effects and naming the condition. Nevertheless, CORE-6D does contain a physical dimension that to some extent picks up comorbidities and perhaps reduces the impact of focusing effects. CORE-6D health state descriptions do not explicitly name the underlying condition, although it is apparent that this involves the presence of mental health symptoms. The analysis of valuation data for CORE-6D presented in Chapter 6 indicated that there was no preference interaction between its emotional and physical components.

The role of generic and condition-specific PBMs has been (and still is) an important subject of debate (Dowie, 2002a & 2002b; Brazier & Tsuchiya, 2010; Brazier & Fitzpatrick, 2002; Feeny, 2002; Guyatt, 2002). Ultimately, the choice between a generic and a condition-specific PBM, such as CORE-6D, in a wider healthcare decision-making context is a trade-off between cross-programme comparability and relevance and sensitivity of the selected PBM in capturing HRQoL changes that are important to patients with the specific condition examined. This remains a controversial issue as well as an area for further research that extends beyond the scope of this thesis.

8.5 The role of patients' preferences in the economic assessment of healthcare interventions

Utility values may be elicited from various groups of stakeholders, including patients, their carers, health professionals and the general public. Selection of one stakeholder group over another may have significant implications in the

estimation of cost-utility of healthcare interventions, as evidence suggests that there are considerable discrepancies in the values obtained from different stakeholders. The current trend, in line with recommendations by advisory and regulatory bodies such as the US Panel on Cost-Effectiveness in Health and Medicine (Russell et al., 1996) and NICE (National Institute for Health and Care Excellence, 2013), is to elicit utility values from a random sample of the general public. This has been the approach at the valuation of the generic PBMs EQ-5D (Dolan et al., 1996), SF-6D (Brazier et al., 2002), and HUI-3 (Feeny et al., 2002), and several condition-specific ones (for example Brazier et al., 2005b & 2008; Yang et al., 2009 & 2011) including the CORE-6D. The main argument for elicitation of preferences from members of the general public is that since health care programmes are funded using society's resources (e.g. through taxation), it is society's preferences that should be taken into account when allocating resources (Gold et al., 1996).

On the other hand, it could be argued that members of the general public have not personally experienced the HRQoL of the health states being valued and therefore they may be lacking the ability to imagine hypothetical health states and to take into account future adaptation to ill health (Brazier et al., 2007; Rowen & Brazier, 2011; Torrance, 1986). A proposed solution to this problem is to provide more information on respondents about the health states being valued, including future adaptation to impaired health, with preliminary evidence indicating that members of the public may change their values in the light of such information (McTaggart-Cowan, 2011; McTaggart-Cowan et al., 2011 & 2012).

In contrast to members of the general public, patients (or carers, when patients are too unwell to provide their own valuations or when the patient population consists of children who do not understand the valuation tasks) can better appreciate the true implications of living in a particular health state and therefore they may be a more appropriate population for eliciting preferences. Moreover, patients may be better suited to be consulted since it is they who are going to be affected by resource allocation decisions. However, one of the dangers in this case is that patients may intentionally or unintentionally

overstate a reduction in the HRQoL relating to their condition, in an effort to ensure access to new treatments that are expected to improve health, as in this case the scope for improvement will look broader. Another problem when eliciting preferences from patients is that they do not have the experience of living under other conditions of health except theirs, so they cannot make judgements across different disease areas (Brazier et al., 2007; Torrance, 1986). Less often, health professionals' values have been used based on their knowledge and expertise in managing the disease area in question. The downside of using this population group is the potential bias that may be introduced due to conflicts of interest and also due to this group's special age, sex, and socio-economic status (Torrance, 1986).

Overall, current evidence indicates that patients value health states more highly than members of the general public (Brazier et al., 2009). This phenomenon has been observed in several disease areas and suggests that either the public does not understand how valuable life can be for patients (or, as argued above, it cannot consider mechanisms of future adaptation), or that patients consciously or subconsciously overstate their HRQoL (Ubel et al., 2003). Another explanation for the discrepancy in values obtained from different stakeholders is that different stakeholder groups have diverse preferences for various types of clinical outcomes. For example, research in the area of schizophrenia has shown that patients rate the importance of extrapyramidal syndrome, a neurological side effect of antipsychotics, more highly than the rest of the stakeholder groups do; clinicians rate social functioning as more important than patients or family members do; clinicians and family members give higher ratings for vocational functioning compared with patients and the general public (Shumway et al., 2003). Such discrepancies in preferences for types of clinical outcomes provide an alternative explanation for the differences in utility values obtained by different stakeholder groups in schizophrenia (Briggs et al., 2008; Lee et al., 2000; Lenert et al., 2000b).

Ultimately, the selection of the valuing population should be determined by the purpose of the study. For comparisons of alternative treatment options in one

patient population, patients with the condition are probably best judges of their HRQoL and it is their preferences that should count. However, for cost-utility analyses undertaken to inform public-funded health service decisions, the appropriate population is probably members of the general public, who are taxpayers and, therefore, funders of the service (Torrance, 1986). Nevertheless, the discrepancy in utility values obtained from different population groups and the importance of considering patients' views and preferences have been the basis for proposals for further research into an approach that directly integrates patients' values into assessments of clinical and cost effectiveness (Brazier et al., 2005a). Direct valuation of CORE-6D by people with common mental health problems is one of the topics recommended for future research in section 8.6.2.

8.6 Recommendations for future research

The discussion of the methods employed in this thesis for the development of CORE-6D, their advantages and limitations, the applicability of the new PBM and the implications for mental health policy brought up a number of issues that warrant further research.

8.6.1 Further use of Rasch analysis in the derivation of preference-based measures from existing measures

The derivation of CORE-6D from CORE-OM was based mainly on Rasch analysis regarding the development of the health state classification, the selection of plausible health states for valuation and the modelling of utility values for all health states described by the new PBM. The latter relied heavily on the relationship between the Rasch logit values of the health states selected for valuation and the corresponding utility values obtained in the valuation survey. Further research into this relationship would allow more vigorous use of Rasch analysis for the derivation of PBMs from existing measures, either unidimensional or with strong unidimensional components, not only regarding the construction of the health state classification (which appears to be its main use so far), but also in the selection of health states for valuation and the subsequent prediction of utility values for all states described by the PBM.

8.6.2 Larger valuation survey and exploration of the preferences of people with common mental health problems

CORE-6D was valued by a random sample of 225 people living in South Yorkshire. Some of the sample's socio-economic characteristics differed from those of the general UK population. Compared with the participants in the valuation of EQ-5D (3,337 respondents) and, to lesser extent, of SF-6D (611 respondents), valuation of CORE-6D used a smaller number of respondents due to funding constraints. In the future it may be worth undertaking re-valuation of the measure with a larger, more representative sample of the general UK population, in order to refine the existing valuation results. Also, a valuation survey of people with common mental health problems will add insight on how this population experiences and values their symptoms and explore discrepancies between the preferences of this patient group and the general population. Ultimately, preferences of people with common mental health problems on HRQoL aspects captured by CORE-6D may play a more active role in the economic assessment of mental health services.

8.6.3 Further validation and testing of the applicability and performance of CORE-6D in a wide range of mental health conditions

As described in Chapter 7, testing of the psychometric properties of CORE-6D and comparison to generic PBMs was characterised by several limitations: analyses were confined to 3 datasets, due to unavailability of other datasets that contained both CORE-OM and other generic measures. Only one of these datasets included a purely 'clinical' population. The known groups validity of CORE-6D and generic PBMs for different levels of mental symptom severity was primarily tested using CORE-OM severity levels, due to lack of another available CSM in the datasets. Similarly, the convergent validity of CORE-6D and generic PBMs was mainly tested against CORE-OM due to lack of other available CSMs. Comparison of CORE-6D with direct (rather than mapped) values of EQ-5D was limited to one small dataset. Comparisons of CORE-6D with mapped EQ-5D values have been likely affected by the applied mapping function and the applicability of this function to populations with common mental health problems. Considering these limitations, future research needs

to validate the results on the psychometric properties of CORE-6D in larger datasets; compare CORE-6D with EQ-5D values that were obtained directly and not by mapping; test the known groups validity of CORE-6D and generic PBMs against a CSM other than CORE-OM that will serve as an independent indicator of symptom severity; assess the convergent validity of CORE-6D and generic PBMs using an alternative CSM and not CORE-OM; and, most importantly, explore the suitability of CORE-6D in a wide range of common mental health problems, including unipolar depression, GAD, panic disorder, OCD, and various types of phobias.

8.6.4 Mapping CORE-10 and CORE-5 onto CORE-6D

The CORE system includes a number of shorter forms of CORE-OM, among which the brief forms CORE-10 and CORE-5 are quite widely used in routine practice for session-by-session monitoring. CORE-6D shares two items with each of the two brief CORE system forms, so it is not possible to obtain CORE-6D values from responses to CORE-10 or CORE-5. Nevertheless, the statistical relationship between each of the two CORE brief forms and CORE-6D can be established by mapping; the resulting algorithms will enable estimation of CORE-6D-based QALYs in datasets that include CORE-10 or CORE-5, but not the full 34-item CORE-OM.

8.6.5 Potential use of CORE-6D as an independent measure

The current proposed use of CORE-6D in a study presupposes use of the CORE-OM, given that CORE-6D items are embedded in the longer 34-item measure. This means that if CORE-OM is not used in a study for outcome measurement, it is not possible to estimate QALYs from CORE-6D. Independent use of CORE-6D has theoretically the advantage that it allows estimation of QALYs with lower burden. However, before CORE-6D is proposed as an independent measure, future research should ensure that the rate and level of responses to the 6 CORE-6D items are not affected by the presence of the rest 28 CORE-OM items, i.e. responses would be the same regardless of whether CORE-6D is embedded in CORE-OM or forms an independent measure.

8.6.6 Use of CORE-6D outside the UK

The translation and validation of CORE-OM in several other languages and countries enables the use of CORE-6D as a PBM in economic evaluations undertaken in mental health settings outside the UK. However, before CORE-6D can be used in other counties, and since its development was based on analysis of UK datasets, the health state descriptive system of CORE-6D (i.e. its items and response levels) needs to be validated by Rasch analysis in other non-UK mental health populations. There is a possibility that different items and health state profiles emerge from such an exercise depending on patients' experiences of mental disease in each particular country, which may indicate the need to derive different health state classifications from CORE-OM for different countries. In any case, whether CORE-6D is validated or a new health state descriptive system is derived as a result of this process, separate valuation surveys need to be conducted in each country, to reflect the preferences of the country's general population.

8.6.7 Development of a 'generic' preference-based measure for all mental disorders

CORE-6D is a PBM relevant to people with common mental health problems as the original CORE-OM has been designed with a focus on this patient population. CORE-6D is probably not appropriate for use in severe, psychotic disorders, as it lacks items that can capture psychotic symptoms. This means that CORE-6D cannot be used as a generic mental health PBM. However, development and use of a generic PBM that is relevant to the full spectrum of mental disorders, covering both common mental health problems and psychotic disorders such as schizophrenia, bipolar disorder and personality disorders would allow wider cost effectiveness comparisons in the area of mental health. *De novo* development of such a measure seems to be the only option, as no generic mental health CSM that is relevant across all areas of mental illness and appropriate for the derivation of a generic mental health PBM is currently available. The new PBM can be developed using as the basis the 7 HRQoL themes that were identified to be most relevant to people with mental disorders (Brazier et al., 2014; Connell et al., 2012).

8.7 Conclusion

Despite the preference of advisory bodies worldwide on the use of generic PBMs for the estimation of QALYs in economic evaluation of healthcare technologies and programmes, these may not be available or appropriate in some conditions and patient populations. In such instances, condition-specific PBMs may prove to be more relevant and sensitive to HRQoL changes. One disease area where generic measures seem to be inappropriate to use is mental health, where the generic measures' inability to capture important qualitative aspects of HRQoL combined with their non-acceptability by patients and healthcare professionals pointed to the need for the construction of a new condition-specific PBM. This need was partly fulfilled with the development of CORE-6D, a PBM derived from the CORE-OM; the latter is a measure routinely used in the NHS that has been designed for the evaluation of psychological services for people with common mental health problems. CORE-6D is a 2-dimensional PBM, capturing emotional and physical symptoms, that has been validated for use in primary and secondary mental healthcare settings. Analyses of its validity and responsiveness in populations with depression and/or anxiety indicated a promising psychometric performance and demonstrated little loss of information compared with the CORE-OM, although further research needs to validate these results and expand analyses to other populations with common mental health problems. Following validation, CORE-6D may be appropriate to use in cost-utility analyses of interventions for common mental health problems, in situations where EQ-5D is not available or has been shown to be inappropriate. The advantages of the use of CORE-6D (and any other condition-specific PBM) over generic PBMs need to be traded-off against compromises in cross-programme comparability, when decisions affecting the wider allocation of healthcare resources are involved, though CORE-6D is less focused on condition-specific problems than many CSMs. The new methods developed for the derivation of CORE-6D from CORE-OM are appropriate for the derivation of new PBMs from instruments with no clear multidimensional structure and/or high correlations across their items, and contribute to the pool of existing methodologies for the derivation of health state classifications from longer measures. These need further exploration, but provide a useful development.

Chapter 9. Appendices

Appendix 1. The structure of the generic preference-based measures EQ-5D, SF-6D and HUI-3

Table A1. The structure of the 3-level response version of EQ-5D
[available on www.euroqol.org (Accessed 23 April 2010)]

Dimension	Level	Statement
Mobility	1	I have no problems in walking about
	2	I have some problems in walking about
	3	I am confined to bed
Self-care	1	I have no problems with self-care
	2	I have some problems washing or dressing myself
	3	I am unable to wash or dress myself
Usual activities *(e.g. work, study, housework, family or leisure activities)*	1	I have no problems with performing my usual activities
	2	I have some problems with performing my usual activities
	3	I am unable to perform my usual activities
Pain/Discomfort	1	I have no pain or discomfort
	2	I have moderate pain or discomfort
	3	I have extreme pain or discomfort
Anxiety/Depression	1	I am not anxious or depressed
	2	I am moderately anxious or depressed
	3	I am extremely anxious or depressed

Table A2. The structure of SF-6D (Brazier et al., 2002)

Dimension	Level	Statement
Physical functioning	1	Your health does not limit you in vigorous activities
	2	Your health limits you a little in vigorous activities
	3	Your health limits you a little in moderate activities
	4	Your health limits you a lot in moderate activities
	5	Your health limits you a little in bathing and dressing
	6	Your health limits you a lot in bathing and dressing
Role limitations	1	You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems
	2	You are limited in the kind of work or other activities as a result of your physical health
	3	You accomplish less than you would like as a result of emotional problems
	4	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems
Social functioning	1	Your health limits your social activities none of the time
	2	Your health limits your social activities a little of the time
	3	Your health limits your social activities some of the time
	4	Your health limits your social activities most of the time
	5	Your health limits your social activities all of the time
Bodily pain	1	You have no pain
	2	You have pain but it does not interfere with your normal work (both outside the home and housework)
	3	You have pain that interferes with your normal work (both outside the home and housework) a little bit
	4	You have pain that interferes with your normal work (both outside the home and housework) moderately
	5	You have pain that interferes with your normal work (both outside the home and housework) quite a bit
	6	You have pain that interferes with your normal work (both outside the home and housework) extremely
Mental health	1	You feel tense or downhearted and low none of the time
	2	You feel tense or downhearted and low a little of the time
	3	You feel tense or downhearted and low some of the time
	4	You feel tense or downhearted and low most of the time
	5	You feel tense or downhearted and low all of the time
Vitality	1	You have a lot of energy all of the time
	2	You have a lot of energy most of the time
	3	You have a lot of energy some of the time
	4	You have a lot of energy a little of the time
	5	You have a lot of energy none of the time

Table A3. The structure of HUI-3 (Feeny et al., 2002)

Attribute	Level	Statement
Vision	1	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street without glasses or contact lenses
	2	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses or contact lenses
	3	Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses or contact lenses
	4	Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses
	5	Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses or contact lenses
	6	Unable to see at all
Hearing	1	Able to hear what is said in a group conversation with at least three other people, without a hearing aid
	2	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people
	3	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid
	4	Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	5	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	6	Unable to hear at all
Speech	1	Able to be understood completely when speaking with strangers or friends
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well
	3	Able to be understood partially when speaking with strangers or people who know me well
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well
	5	Unable to be understood when speaking to other people (or unable to speak at all)
Ambulation	1	Able to walk around the neighbourhood without difficulty, and without walking equipment
	2	Able to walk around the neighbourhood with difficulty, but does not require walking equipment or the help of another person
	3	Able to walk around the neighbourhood with walking equipment, but without the help of another person
	4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood
	5	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood
	6	Cannot walk at all

Dexterity	1	Full use of two hands and ten fingers
	2	Limitations in the use of hands or fingers, but does not require special tools or help of another person
	3	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person)
	4	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools)
	5	Limitations in the use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools)
	6	Limitations in the use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools)
Emotion	1	Happy and interested in life
	2	Somewhat happy
	3	Somewhat unhappy
	4	Very unhappy
	5	So unhappy that life is not worthwhile
Cognition	1	Able to remember most things, think clearly and solve day to day problems
	2	Able to remember most things, but have a little difficulty when trying to think and solve day to day problems
	3	Somewhat forgetful but able to think clearly and solve day to day problems
	4	Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems
	5	Very forgetful, and have great difficulty when trying to think or solve day to day problems
	6	Unable to remember anything at all, and unable to think or solve day to day problems
Pain	1	Free of pain and discomfort
	2	Mild to moderate pain that prevents no activities
	3	Moderate pain that prevents a few activities
	4	Moderate to severe pain that prevents some activities
	5	Severe pain that prevents most activities

Appendix 2. Search strategy used for the identification of reviews assessing the use of generic preference-based measures and condition-specific outcome measures in mental health research and practice

Databases	**Embase <1980 to 2012 Week 50>, Ovid MEDLINE(R) <1946 to November Week 3 2012>, PsycINFO <1806 to December Week 2 2012>**	
Interface	OvidSP	
Date of search	20 December 2012	
Search Strategy:		
No	**Search terms**	**Hits**
1	exp mental health/ or exp psychiatry/ or exp mental disease/	1,695,768
2	1 use emez	1,524,932
3	exp Mental Health/ or exp Psychiatry/ or exp Mental Disorders/	2,935,053
4	3 use mesz	959,496
5	3 use psyh	450,625
6	2 or 4 or 5	2,935,053
7	mental health.ti,ab.	249,188
8	mental* ill*.ti,ab.	77,699
9	mental* ill-health.ti,ab.	1,090
10	psychiatr*.ti,ab.	549,795
11	mental* disorder*.ti,ab.	79,590
12	or/6-11	3,180,951
13	exp "quality of life"/ or exp treatment outcome/ or exp psychologic assessment/ or exp health survey/ or exp psychotherapy/ or exp outcomes research/ or exp outcome assessment/	2,684,683
14	13 use emez	1,341,205
15	exp "Quality of Life"/ or exp Treatment Outcome/ or exp Health Surveys/ or exp Psychotherapy/ or exp "Outcome Assessment (Health Care)"/	2,672,645
16	15 use mesz	1,137,873
17	exp "Quality of Life"/ or exp Treatment Outcomes/ or exp Psychological Assessment/ or Surveys/ or exp Psychotherapy/	955,084
18	17 use psyh	239,680
19	or/14,16,18	2,718,758
20	treatment effectiveness evaluation.mp.	14,102
21	health status indicator*.ti,ab	696
22	health outcome*.ti,ab	45,694
23	((quality adj1 life) or qol or hrqol or hrql or hql or hqol or qaly or quality adjusted life or qwb).ti,ab.	88,307
24	(quality adj1 (wellbeing or well being)).ti,ab.	348
25	((measur* or assess* or scor* or index* or indices or scal* or	383,540

	monitor*) adj2 outcome*).ti,ab.	
26	((improv* or measur*) adj1 (productivity or performance)).ti,ab.	47,943
27	((output or price) adj (index* or indices)).ti,ab.	1,375
28	outcome measure*.ti,ab.	296,694
29	((utilit* or preference) adj1 (based or index* or indices or measure* or valu* or scor* or weigh*)).ti,ab.	9,897
30	(euroqol* or euro qol* or eq5d* or eq 5d*).ti,ab.	8,845
31	(hui or hui1 or hui2 or hui3).ti,ab.	2,088
32	(sf36 or sf 36 or short form 36 or shortform 36 or sf thirtysix or sf thirty six or shortform thirtysix or shortform thirty six or short form thirtysix or short form thirty six).ti,ab.	35,459
33	(sf6 or sf 6 or short form 6 or shortform 6 or sf six or sfsix or shortform six or short form six).ti,ab.	2,288
34	(sf12 or sf 12 or short form 12 or shortform 12 or sf twelve or sftwelve or shortform twelve or short form twelve).ti,ab.	5,911
35	(sf16 or sf 16 or short form 16 or shortform 16 or sf sixteen or sfsixteen or shortform sixteen or short form sixteen).ti,ab.	46
36	(sf20 or sf 20 or short form 20 or shortform 20 or sf twenty or sftwenty or shortform twenty or short form twenty).ti,ab.	661
37	or/19-36	3,054,371
38	12 and 37	532,216
39	limit 38 to (human and english language and "reviews (maximizes specificity)" and yr="2002 -Current")	8,658

Database	**HMIC Health Management Information Consortium <1979 to November 2012>**	
Interface	OvidSP	
Date of search	20 December 2012	
Search Strategy:		
No	**Search terms**	**Hits**
1	exp Mental health/ or exp Psychiatry/ or exp Mental illness/	11,660
2	mental health.ti,ab.	15,532
3	mental* ill*.ti,ab.	4,335
4	mental* ill-health.ti,ab.	174
5	psychiatr*.ti,ab.	8,130
6	mental* disorder*.ti,ab.	1,378
7	or/1-6	26,474
8	exp "Quality of life"/ or exp Patient outcome/ or exp Health surveys/ or exp Psychotherapy/ or exp Outcome measurement/	9,297
9	treatment outcome.mp.	114
10	outcome research.mp.	51
11	health status indicator*.ti,ab.	13
12	health outcome*.ti,ab.	1,973
13	((quality adj1 life) or qol or hrqol or hrql or hql or hqol or qaly or quality adjusted life or qwb).ti,ab.	1,115
14	(quality adj1 (wellbeing or well being)).ti,ab.	8

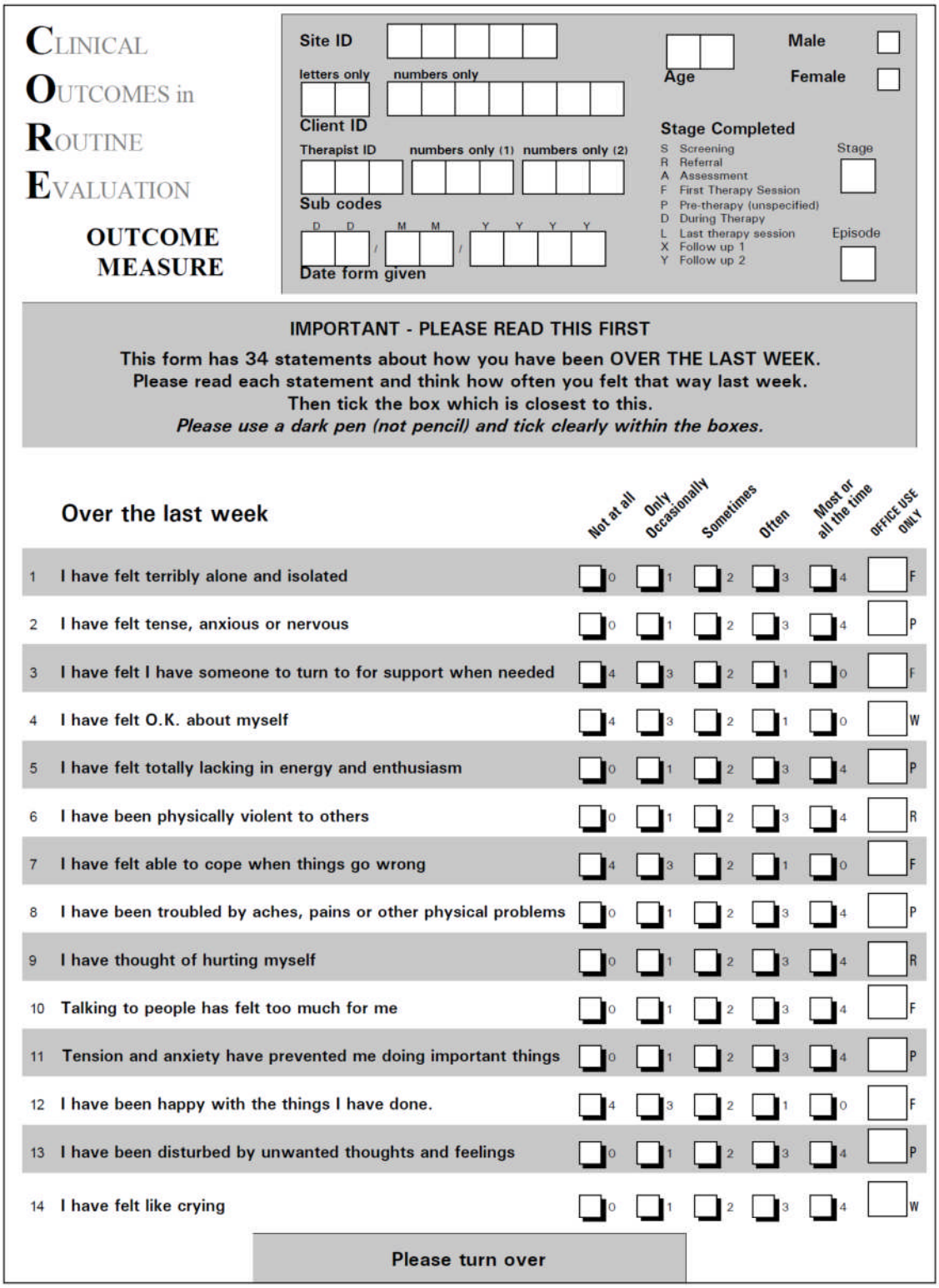
15	((measur* or assess* or scor* or index* or indices or scal* or monitor*) adj2 outcome*).ti,ab.	7,475
16	((improv* or measur*) adj1 (productivity or performance)).ti,ab.	1,159
17	((output or price) adj (index* or indices)).ti,ab.	68
18	outcome measure*.ti,ab.	6,119
19	((utilit* or preference) adj1 (based or index* or indices or measure* or valu* or scor* or weigh*)).ti,ab.	176
20	(euroqol* or euro qol* or eq5d* or eq 5d*).ti,ab.	250
21	(hui or hui1 or hui2 or hui3).ti,ab.	20
22	(sf36 or sf 36 or short form 36 or shortform 36 or sf thirtysix or sf thirty six or shortform thirtysix or shortform thirty six or short form thirtysix or short form thirty six).ti,ab.	376
23	(sf6 or sf 6 or short form 6 or shortform 6 or sf six or sfsix or shortform six or short form six).ti,ab.	2
24	(sf12 or sf 12 or short form 12 or shortform 12 or sf twelve or sftwelve or shortform twelve or short form twelve).ti,ab.	68
25	(sf16 or sf 16 or short form 16 or shortform 16 or sf sixteen or sfsixteen or shortform sixteen or short form sixteen).ti,ab.	0
26	(sf20 or sf 20 or short form 20 or shortform 20 or sf twenty or sftwenty or shortform twenty or short form twenty).ti,ab.	2
27	or/8-26	18,837
28	7 and 27	2,222
29	limit 28 to (yr="2002 -Current" and english)	1,306

Databases	**Cochrane Database of Systematic Reviews (CDSR), Cochrane Methodology Register, Health Technology Assessment (HTA) database, Database of Abstracts of Reviews of Effects (DARE)**	
Interface	Wiley	
Date of search	20 December 2012	
Search Strategy:		
No	**Search terms**	**Hits**
1	MeSH descriptor: [Mental Health] explode all trees	487
2	MeSH descriptor: [Psychiatry] explode all trees	399
3	MeSH descriptor: [Mental Disorders] explode all trees	37,889
4	"mental health":ti,ab	3,123
5	mental* next ill*:ti,ab	1,030
6	mental* next ill-health:ti,ab	4
7	psychiatr*:ti,ab	6,321
8	mental* next disorder*:ti,ab	1,965
9	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8	43,485
10	MeSH descriptor: [Quality of Life] explode all trees	12,121
11	MeSH descriptor: [Health Surveys] explode all trees	20,840
12	MeSH descriptor: [Psychotherapy] explode all trees	13,097
13	MeSH descriptor: [Outcome Assessment (Health Care)] explode all trees	82,442

14	treatment next outcome*:ti,ab	29,647
15	"outcome research":ti,ab	64
16	health next status next indicator*:ti,ab	44
17	health next outcome*:ti,ab	1,909
18	((quality near/2 life) or qol or hrqol or hrql or hql or hqol or qaly or (quality next adjusted next life) or qwb):ti,ab	19,089
19	(quality near/2 (wellbeing or well next being)):ti,ab	146
20	((measur* or assess* or scor* or index* or indices or scal* or monitor*) near/3 outcome*):ti,ab	36,554
21	((improv* or measur*) near/2 (productivity or performance)):ti,ab	3,707
22	((output or price) next (index* or indices)):ti,ab	14
23	outcome next measure*:ti,ab	28,517
24	((utilit* or preference) next (based or index* or indices or measure* or valu* or scor* or weigh*)):ti,ab	401
25	(euroqol* or euro qol* or eq5d* or eq 5d*):ti,ab	710
26	(hui or hui1 or hui2 or hui3):ti,ab	71
27	(sf36 or sf 36 or "short form 36" or "shortform 36" or "sf thirtysix" or "sf thirty six" or "shortform thirtysix" or "shortform thirty six" or "short form thirtysix" or "short form thirty six"):ti,ab	2,744
28	(sf6 or "sf 6" or "short form 6" or "shortform 6" or "sf six" or sfsix or "shortform six" or "short form six"):ti,ab	73
29	(sf12 or "sf 12" or "short form 12" or "shortform 12" or "sf twelve" or sftwelve or "shortform twelve" or "short form twelve"):ti,ab	557
30	(sf16 or "sf 16" or "short form 16" or "shortform 16" or "sf sixteen" or sfsixteen or "shortform sixteen" or "short form sixteen"):ti,ab	5
31	(sf20 or "sf 20" or "short form 20" or "shortform 20" or "sf twenty" or sftwenty or "shortform twenty" or "short form twenty"):ti,ab	56
32	#10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24 or #25 or #26 or #27 or #28 or #29 or #30 or #31	143,558
33	#9 and #32	20,231
34	#34 from 2002, in Cochrane Reviews (Reviews only), Other Reviews, Methods Studies, Technology Assessments and Cochrane Groups	1,651

Appendix 3. The CORE Outcome Measure form

Reproduced with kind permission of the CORE System Trust



CLINICAL
OUTCOMES in
ROUTINE
EVALUATION

OUTCOME MEASURE

Site ID

letters only | numbers only

Client ID

Therapist ID | numbers only (1) | numbers only (2)

Sub codes

D D / M M / Y Y Y Y

Date form given

Age

Male

Female

Stage Completed

S Screening
R Referral
A Assessment
F First Therapy Session
P Pre-therapy (unspecified)
D During Therapy
L Last therapy session
X Follow up 1
Y Follow up 2

Stage

Episode

IMPORTANT - PLEASE READ THIS FIRST

This form has 34 statements about how you have been OVER THE LAST WEEK.
Please read each statement and think how often you felt that way last week.
Then tick the box which is closest to this.
Please use a dark pen (not pencil) and tick clearly within the boxes.

	Over the last week	Not at all	Only Occasionally	Sometimes	Often	Most or all the time	OFFICE USE ONLY
1	I have felt terribly alone and isolated	0	1	2	3	4	F
2	I have felt tense, anxious or nervous	0	1	2	3	4	P
3	I have felt I have someone to turn to for support when needed	4	3	2	1	0	F
4	I have felt O.K. about myself	4	3	2	1	0	W
5	I have felt totally lacking in energy and enthusiasm	0	1	2	3	4	P
6	I have been physically violent to others	0	1	2	3	4	R
7	I have felt able to cope when things go wrong	4	3	2	1	0	F
8	I have been troubled by aches, pains or other physical problems	0	1	2	3	4	P
9	I have thought of hurting myself	0	1	2	3	4	R
10	Talking to people has felt too much for me	0	1	2	3	4	F
11	Tension and anxiety have prevented me doing important things	0	1	2	3	4	P
12	I have been happy with the things I have done.	4	3	2	1	0	F
13	I have been disturbed by unwanted thoughts and feelings	0	1	2	3	4	P
14	I have felt like crying	0	1	2	3	4	W

Please turn over

Over the last week

		Not at all	Only Occasionally	Sometimes	Often	Most or all the time	OFFICE USE ONLY
15	I have felt panic or terror	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
16	I made plans to end my life	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ R
17	I have felt overwhelmed by my problems	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ W
18	I have had difficulty getting to sleep or staying asleep	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
19	I have felt warmth or affection for someone	☐ 4	☐ 3	☐ 2	☐ 1	☐ 0	☐ F
20	My problems have been impossible to put to one side	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
21	I have been able to do most things I needed to	☐ 4	☐ 3	☐ 2	☐ 1	☐ 0	☐ F
22	I have threatened or intimidated another person	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ R
23	I have felt despairing or hopeless	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
24	I have thought it would be better if I were dead	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ R
25	I have felt criticised by other people	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ F
26	I have thought I have no friends	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ F
27	I have felt unhappy	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
28	Unwanted images or memories have been distressing me	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
29	I have been irritable when with other people	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ F
30	I have thought I am to blame for my problems and difficulties	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ P
31	I have felt optimistic about my future	☐ 4	☐ 3	☐ 2	☐ 1	☐ 0	☐ W
32	I have achieved the things I wanted to	☐ 4	☐ 3	☐ 2	☐ 1	☐ 0	☐ F
33	I have felt humiliated or shamed by other people	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ F
34	I have hurt myself physically or taken dangerous risks with my health	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ R

THANK YOU FOR YOUR TIME IN COMPLETING THIS QUESTIONNAIRE

	(W)	(P)	(F)	(R)	All items	All minus R
Total Scores	☐	☐	☐	☐ →	☐ →	☐
Mean Scores (Total score for each dimension divided by number of items completed in that dimension)	☐	☐	☐	☐	☐	☐

Appendix 4. Search strategy used for the identification of studies reporting methods for the derivation of health state descriptions from existing non-preference-based measures

Databases	**Embase <1980 to 2012 Week 50>, HMIC Health Management Information Consortium <1979 to November 2012>, Ovid MEDLINE(R) <1946 to November Week 3 2012>, PsycINFO <1806 to December Week 3 2012**	
Interface	OvidSP	
Date of search	21 December 2012	
Search Strategy:		
No	**Search terms**	**Hits**
1	exp "quality of life"/	353,400
2	((quality adj1 life) or qol or hrqol or hrql or hql or hqol or qaly or quality adjusted life or quality-adjusted life or qwb).ti,ab.	89,440
3	(quality adj1 (wellbeing or well being)).ti,ab.	356
4	1 or 2 or 3	369,409
5	((utilit* or preference) adj1 (based or index* or indices or measure* or valu* or scor* or weigh*)).ti,ab.	10,075
6	((health state* or health-state*) adj1 (descri* or classification* or valu*)).ti,ab.	1,103
7	(condition-specific or condition specific).ti,ab.	2,432
8	5 or 6 or 7	12,969
9	(transform* or translat* or transfer* or develop* or conver* or map* or deriv*).ti,ab.	9,836,495
10	4 and 8 and 9	2,607
11	limit 10 to english language [Limit not valid in HMIC; records were retained]	2,564
12	limit 11 to human [Limit not valid in HMIC; records were retained]	2,341
13	remove duplicates from 12	1,397

Databases	**Cochrane Database of Systematic Reviews (CDSR), Cochrane Methodology Register, Health Technology Assessment (HTA) database, Database of Abstracts of Reviews of Effects (DARE)**	
Interface	Wiley	
Date of search	21 December 2012	
Search Strategy:		
No	**Search terms**	**Hits**
1	MeSH descriptor: [Quality of Life] explode all trees	12,121
2	((quality near/2 life) or qol or hrqol or hrql or hql or hqol or qaly or (quality next adjusted next life) or (quality-adjusted next life) or qwb):ti,ab	19,089
3	(quality near/2 (wellbeing or well next being)):ti,ab	146
4	#1 or #2 or #3	22,596
5	((utilit* or preference) next (based or index* or indices or measure* or valu* or scor* or weigh*)):ti,ab	401
6	((health next state* or “health-state”) next (descri* or classification* or valu*)):ti,ab	25
7	(“condition-specific” or “condition specific”):ti,ab	2,827
8	#5 or #6 or #7	3,233
9	(transform* or translat* or develop* or transfer* or conver* or map* or deriv*):ti,ab	81,349
10	#4 and #8 and #9	136

Appendix 5. Results of Principal Components Analysis on CORE-OM data in study samples [N750a] and [N750b]

PCA analysis in [N750a]

Table A4. Significant components of CORE-OM identified by Principal Components Analysis in study sample [N750a]

Component	PCA: Initial Eigenvalues			Horn's parallel analysis: Significant mean eigenvalues (SD)	
	Total	% of Variance	Cumulative %		
1	**11.26**	33.2	33.2	**1.42**	(0.03)
2	**2.14**	6.3	39.5	**1.37**	(0.02)
3	**1.83**	5.4	44.9	**1.33**	(0.02)
4	**1.48**	4.3	49.2	**1.30**	(0.02)
5	**1.30**	3.8	53.0	**1.27**	(0.02)
6	**1.08**	3.2	56.2	1.24	(0.02)

Significant eigenvalue levels identified using each approach are provided in bold; SD = standard deviation

Figure A1. Screeplot of Principal Components Analysis in [N750a]

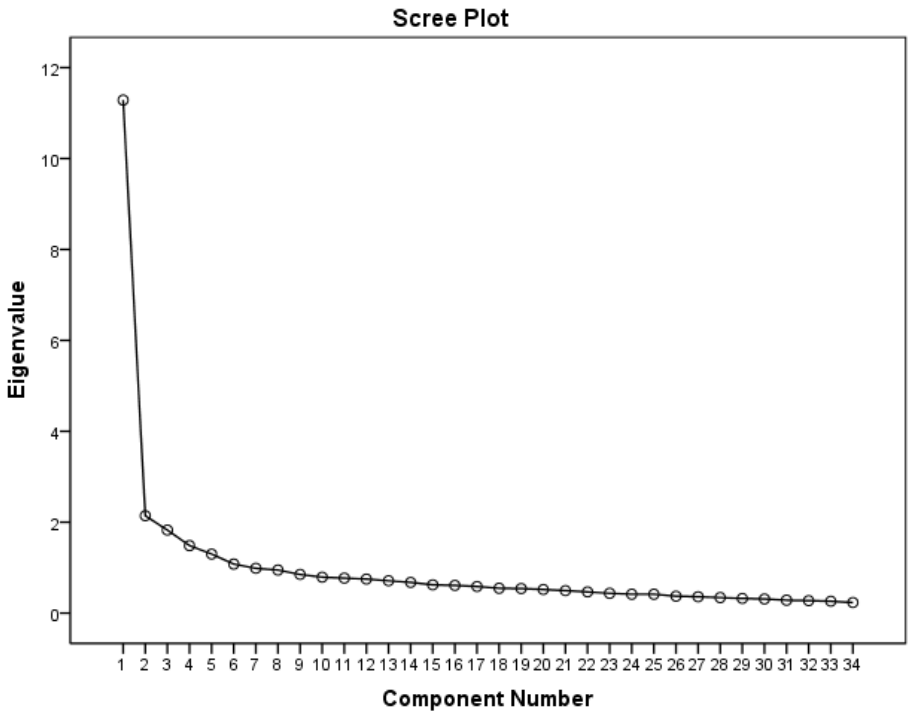


Table A5. Findings of Principal Components Analysis on CORE-OM data in study sample [N750a]. Orthogonal rotation - rotated component matrix

CORE-OM items	Components					
	1	2	3	4	5	6
1. [terribly alone and isolated]	**.60**	.29	.30	.19	-.02	.08
2. [tense, anxious or nervous]	**.61**	.23	.08	.00	.16	.34
3. [somebody to turn to for support]	-.01	.33	**.64**	.02	-.10	.12
4. [felt ok about myself]	**.40**	**.59**	.25	.14	.01	.03
5. [totally lacking in energy and enthusiasm]	**.46**	.36	.11	.03	.12	.33
6. [physically violent to others]	.06	.02	-.05	.18	**.78**	.04
7. [able to cope when things go wrong]	**.40**	**.55**	.07	.13	.16	-.08
8. [aches, pains, physical problems]	.11	-.01	.03	.04	.02	**.82**
9. [thought of hurting myself]	.23	.12	.09	**.84**	.07	.03
10. [talking to people has felt too much]	.25	.24	**.42**	.13	.02	.33
11. [tension/anxiety prevented doing things]	**.52**	.37	-.04	.08	.20	.35
12. [happy with the things I've done]	.31	**.66**	.21	.06	.12	-.06
13. [disturbed by unwanted thoughts/feelings]	**.65**	.02	.16	.18	.01	.12
14. [felt like crying]	**.62**	.14	.25	.15	.03	-.06
15. [felt panic or terror]	**.66**	.12	-.11	.19	.08	.27
16. [made plans to end my life]	.16	.12	.05	**.82**	.03	.07
17. [overwhelmed by my problems]	**.70**	.33	.16	.11	.15	.07
18. [difficulty of getting to sleep/staying asleep]	**.47**	.07	.26	.04	-.03	.38
19. [felt warmth or affection for someone]	-.09	**.60**	.18	.08	-.11	-.03
20. [problems impossible to put to one side]	**.69**	.23	.16	.03	.13	.02
21. [able to do most things I needed to]	.35	**.62**	-.05	.03	.18	.08
22. [threatened or intimidated another person]	.06	.04	.26	.06	**.73**	.01
23. [felt despairing or hopeless]	**.60**	.38	.25	.23	.12	.10
24. [thought it would be better if I were dead]	.28	.22	.25	**.71**	.10	-.01
25. [felt criticised by other people]	.32	.13	**.57**	.10	.34	.04
26. [thought I have no friends]	.28	.22	**.60**	.19	.07	.03
27. [felt unhappy]	**.64**	.35	.31	.13	-.01	-.03
28. [unwanted images/memories distressing]	**.68**	-.06	.28	.18	.00	-.05
29. [irritable when with other people]	.35	.08	**.49**	.06	.36	.10
30. [I am to blame for problems & difficulties]	.31	.19	**.43**	.15	.10	-.15
31. [felt optimistic about my future]	.05	**.65**	.15	.10	-.10	.14
32. [achieved the things I wanted to]	.30	**.64**	.16	.14	.11	.10
33. [felt humiliated or shamed by other people]	.36	.04	**.50**	.21	.33	-.00
34. [hurt myself physically/risks with health]	.02	.07	.15	**.57**	.39	.04

Rotation method Varimax with Kaiser normalisation. Loadings ≥ $\left|0.40\right|$ are shown in bold; loadings on the component that was found to be non-significant by Horn's analysis have been shaded in grey.

Table A6. Findings of Principal Components Analysis on CORE-OM data in study sample [N750a]. Oblique rotation - pattern matrix

CORE-OM items	Components					
	1	2	3	4	5	6
1. [terribly alone and isolated]	**.61**	.11	.17	.07	-.14	-.00
2. [tense, anxious or nervous]	**.63**	.06	-.04	-.11	.08	.26
3. [somebody to turn to for support]	-.26	.30	**.73**	-.07	-.15	.14
4. [felt ok about myself]	.26	**.54**	.14	.03	-.07	-.06
5. [totally lacking in energy and enthusiasm]	.39	.25	.02	-.07	.05	.25
6. [physically violent to others]	-.13	-.02	-.12	.12	**.84**	-.00
7. [able to cope when things go wrong]	.29	**.53**	-.08	.02	.10	-.18
8. [aches, pains, physical problems]	-.02	-.11	.10	.05	.00	**.85**
9. [thought of hurting myself]	.10	-.00	-.06	**.87**	-.01	.01
10. [talking to people has felt too much]	.08	.12	**.43**	.05	-.06	.32
11. [tension/anxiety prevented doing things]	**.46**	.26	-.18	-.01	.13	.26
12. [happy with the things I've done]	.13	**.67**	.10	-.06	.06	-.15
13. [disturbed by unwanted thoughts/feelings]	**.78**	-.21	.03	.08	-.11	.05
14. [felt like crying]	**.71**	-.05	.10	.02	-.09	-.15
15. [felt panic or terror]	**.78**	-.07	-.29	.12	-.02	.18
16. [made plans to end my life]	.01	.02	-.08	**.87**	-.04	.06
17. [overwhelmed by my problems]	**.73**	.15	-.02	-.04	.05	-.04
18. [difficulty of getting to sleep/staying asleep]	**.48**	-.12	.22	-.05	-.13	.34
19. [felt warmth or affection for someone]	-.36	**.72**	.17	.06	-.12	-.05
20. [problems impossible to put to one side]	**.79**	.04	-.01	-.12	.03	-.09
21. [able to do most things I needed to]	.21	**.64**	-.20	-.07	.14	-.03
22. [threatened or intimidated another person]	-.16	-.03	.25	-.06	**.77**	-.01
23. [felt despairing or hopeless]	**.55**	.22	.09	.11	.01	.01
24. [thought it would be better if I were dead]	.12	.10	.12	**.68**	.01	-.05
25. [felt criticised by other people]	.16	-.04	**.56**	-.06	.29	.01
26. [thought I have no friends]	.12	.07	**.59**	.07	-.02	.01
27. [felt unhappy]	**.67**	.17	.16	-.02	-.13	-.12
28. [unwanted images/memories distressing]	**.86**	-.32	.15	.05	-.13	-.12
29. [irritable when with other people]	.23	-.10	**.48**	-.10	.31	.07
30. [I am to blame for problems & difficulties]	.25	.07	.37	.03	.03	-.19
31. [felt optimistic about my future]	-.21	**.73**	.12	.06	-.13	.10
32. [achieved the things I wanted to]	.08	**.63**	.06	.05	.06	.02
33. [felt humiliated or shamed by other people]	.25	-.15	**.46**	.07	.27	-.03
34. [hurt myself physically/risks with health]	-.21	.00	.09	**.56**	.38	.04

Rotation method Promax with Kaiser normalisation. Loadings ≥ |0.40| are shown in bold; loadings on the component that was found to be non-significant by Horn's analysis have been shaded in grey.

Table A7. Findings of Principal Components Analysis of CORE-OM data in study random sub-sample [N750a]. Oblique rotation – component correlation matrix

Component	1	2	3	4	5	6
1	**1.00**	**0.58**	**0.50**	**0.41**	**0.42**	0.27
2	**0.58**	**1.00**	0.38	0.29	0.23	0.22
3	**0.50**	0.38	**1.00**	0.36	0.26	-0.01
4	**0.41**	0.29	0.36	**1.00**	0.28	0.02
5	**0.42**	0.23	0.26	0.28	**1.00**	0.11
6	0.27	0.22	-0.01	0.02	0.11	1.00

Rotation method Promax with Kaiser normalisation. Correlation coefficients ≥ │0.40│ are shown in bold; correlations of the component that was found to be non-significant by Horn's analysis have been shaded in grey.

PCA analysis on [N750b]

Table A8. Significant components of CORE-OM identified by Principal Components Analysis in study sample [N750b]

Component	PCA: Initial Eigenvalues			Horn's parallel analysis: Significant mean eigenvalues (SD)	
	Total	% of Variance	Cumulative %		
1	**11.05**	32.5	32.5	**1.42**	(0.03)
2	**2.15**	6.3	38.8	**1.37**	(0.02)
3	**1.78**	5.2	44.0	**1.33**	(0.02)
4	**1.45**	4.3	48.3	**1.30**	(0.02)
5	**1.22**	3.6	51.9	1.27	(0.02)
6	**1.13**	3.3	55.2	1.24	(0.02)
7	**1.07**	3.2	58.4	1.21	(0.01)

Significant eigenvalue levels identified using each approach are provided in bold; SD = standard deviation

Figure A2. Screeplot of Principal Components Analysis in [N750b]

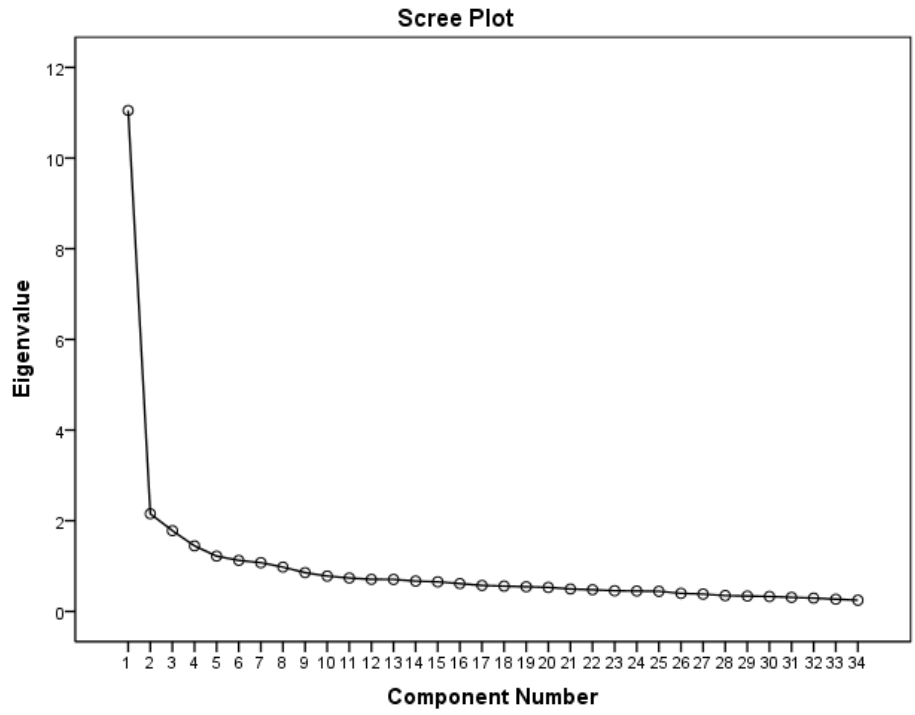


Table A9. Findings of Principal Components Analysis on CORE-OM data in study sample [N750b]. Orthogonal rotation - rotated component matrix

CORE-OM items	Components						
	1	2	3	4	5	6	7
1. [terribly alone and isolated]	**.44**	.39	.28	.16	.24	.24	-.07
2. [tense, anxious or nervous]	**.41**	.26	**.58**	-.02	.15	-.04	.00
3. [somebody to turn to for support]	.08	.23	.03	.08	.21	**.71**	.03
4. [felt ok about myself]	.34	**.58**	.09	.13	.12	.22	.02
5. [totally lacking in energy and enthusiasm]	.37	.34	**.42**	.09	.04	.04	-.02
6. [physically violent to others]	.05	.11	.02	.18	.03	.09	**.79**
7. [able to cope when things go wrong]	.25	**.63**	.25	.15	.04	-.04	.14
8. [aches, pains, physical problems]	.17	-.22	**.59**	.08	-.04	.36	-.13
9. [thought of hurting myself]	.14	.20	.11	**.82**	.09	.07	.10
10. [talking to people has felt too much]	.25	.24	**.47**	.03	.12	.26	.10
11. [tension/anxiety prevented doing things]	.15	**.42**	**.65**	.13	.23	-.05	.14
12. [happy with the things I've done]	.26	**.70**	.08	.07	.16	.17	.11
13. [disturbed by unwanted thoughts/feelings]	**.58**	.00	.24	.27	.18	-.00	-.12
14. [felt like crying]	**.67**	.26	.21	.10	.15	-.02	.05
15. [felt panic or terror]	.24	.22	**.59**	.20	.15	-.11	-.01
16. [made plans to end my life]	.10	.07	.12	**.80**	.09	.01	.07
17. [overwhelmed by my problems]	**.53**	.37	.37	.12	.26	-.02	.02
18. [difficulty of getting to sleep/staying asleep]	**.62**	.20	.07	.13	-.12	.14	.07
19. [felt warmth or affection for someone]	-.01	.20	.02	.12	.02	**.69**	.18
20. [problems impossible to put to one side]	**.60**	.29	.27	.03	.20	-.01	.08
21. [able to do most things I needed to]	.03	**.62**	.36	.17	.08	.01	.23
22. [threatened or intimidated another person]	.06	-.03	.00	.10	.23	.07	**.76**
23. [felt despairing or hopeless]	**.43**	.39	.35	.18	.38	.08	.04
24. [thought it would be better if I were dead]	.28	.17	.17	**.67**	.34	.06	-.02
25. [felt criticised by other people]	.18	.12	.09	.04	**.77**	.07	.14
26. [thought I have no friends]	.16	.15	.14	.15	**.59**	.34	.13
27. [felt unhappy]	**.61**	**.42**	.20	.08	.24	.12	.04
28. [unwanted images/memories distressing]	**.66**	-.01	.11	.28	.15	.01	.06
29. [irritable when with other people]	**.50**	.18	.07	-.01	.27	.12	.34
30. [I am to blame for problems & difficulties]	.27	.37	-.04	.23	**.50**	.04	-.06
31. [felt optimistic about my future]	.24	**.56**	.11	.01	.06	.28	-.18
32. [achieved the things I wanted to]	.08	**.66**	.07	.12	.22	.21	-.01
33. [felt humiliated or shamed by other people]	.11	.11	.21	.14	**.72**	.01	.14
34. [hurt myself physically/risks with health]	.14	.06	-.02	**.61**	.03	.16	.19

Rotation method Varimax with Kaiser normalisation. Loadings ≥ |0.40| are shown in bold; loadings on the 3 components that were found to be non-significant by Horn's analysis have been shaded in grey.

Table A10. Findings of Principal Components Analysis on CORE-OM data in study sample [N750b]. Oblique rotation - pattern matrix

CORE-OM items	Components						
	1	2	3	4	5	6	7
1. [terribly alone and isolated]	.31	.27	.16	.03	.10	.17	-.12
2. [tense, anxious or nervous]	.26	.09	**.57**	-.15	.03	-.09	-.02
3. [somebody to turn to for support]	-.06	.19	-.01	.00	.16	**.69**	-.02
4. [felt ok about myself]	.22	**.61**	-.07	.03	-.06	.15	-.04
5. [totally lacking in energy and enthusiasm]	.26	.25	.38	-.01	-.14	-.02	-.04
6. [physically violent to others]	.06	.05	-.01	.12	-.08	.08	**.79**
7. [able to cope when things go wrong]	.08	**.67**	.13	.07	-.16	-.11	.08
8. [aches, pains, physical problems]	.08	**-.45**	**.76**	.03	-.11	.37	-.11
9. [thought of hurting myself]	.01	.13	.01	**.84**	-.07	.02	.04
10. [talking to people has felt too much]	.09	.10	**.49**	-.09	.07	.22	.08
11. [tension/anxiety prevented doing things]	-.18	.31	**.66**	.01	.12	-.10	.09
12. [happy with the things I've done]	.09	**.77**	-.09	-.04	-.01	.10	.04
13. [disturbed by unwanted thoughts/feelings]	**.63**	-.24	.14	.18	.07	-.07	-.14
14. [felt like crying]	**.73**	.08	.05	-.03	-.03	-.11	.03
15. [felt panic or terror]	.02	.07	**.61**	.11	.05	-.15	-.04
16. [made plans to end my life]	-.02	-.02	.05	**.84**	-.03	-.03	.02
17. [overwhelmed by my problems]	**.43**	.21	.24	-.02	.11	-.11	-.03
18. [difficulty of getting to sleep/staying asleep]	**.78**	.09	-.07	.05	-.35	.08	.07
19. [felt warmth or affection for someone]	-.12	.20	.03	.07	-.07	**.69**	.15
20. [problems impossible to put to one side]	**.60**	.12	.13	-.12	.05	-.09	.05
21. [able to do most things I needed to]	-.26	**.68**	.32	.10	-.08	-.04	.16
22. [threatened or intimidated another person]	.08	-.16	-.03	.02	.22	.05	**.75**
23. [felt despairing or hopeless]	.25	.23	.22	.04	.27	-.01	-.03
24. [thought it would be better if I were dead]	.12	.02	.04	**.62**	.25	-.01	-.09
25. [felt criticised by other people]	-.02	-.07	-.04	-.11	**.90**	.01	.06
26. [thought I have no friends]	-.04	-.03	.06	.02	**.64**	.29	.06
27. [felt unhappy]	**.58**	.28	.01	-.07	.08	.03	-.00
28. [unwanted images/memories distressing]	**.81**	-.26	-.05	.19	.02	-.07	.05
29. [irritable when with other people]	**.56**	.01	-.08	-.16	.17	.06	.32
30. [I am to blame for problems & difficulties]	.12	.30	-.25	.13	**.50**	-.05	-.15
31. [felt optimistic about my future]	.10	**.63**	-.01	-.07	-.08	.22	-.24
32. [achieved the things I wanted to]	-.18	**.77**	-.07	.04	.12	.15	-.09
33. [felt humiliated or shamed by other people]	-.14	-.09	.13	.01	**.83**	-.05	.07
34. [hurt myself physically/risks with health]	.12	-.01	-.10	**.63**	-.09	.13	.16

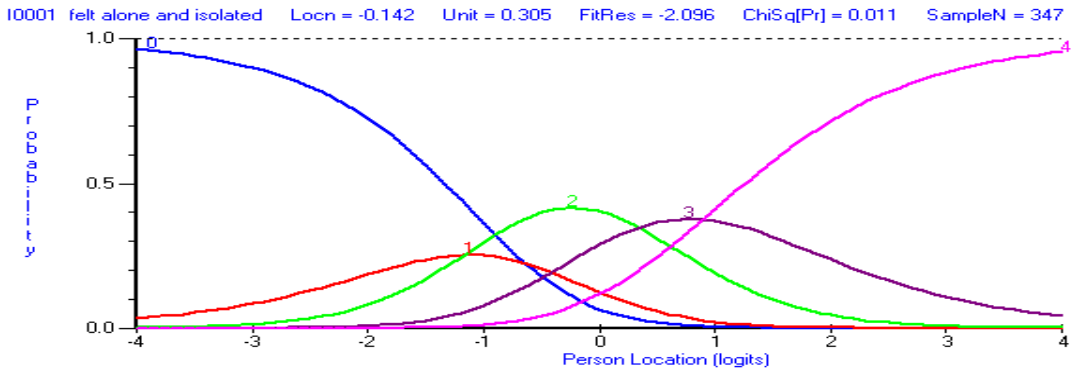
Rotation method Promax with Kaiser normalisation. Loadings ≥ |0.40| are shown in bold; loadings on the 3 components that were found to be non-significant by Horn's analysis have been shaded in grey.

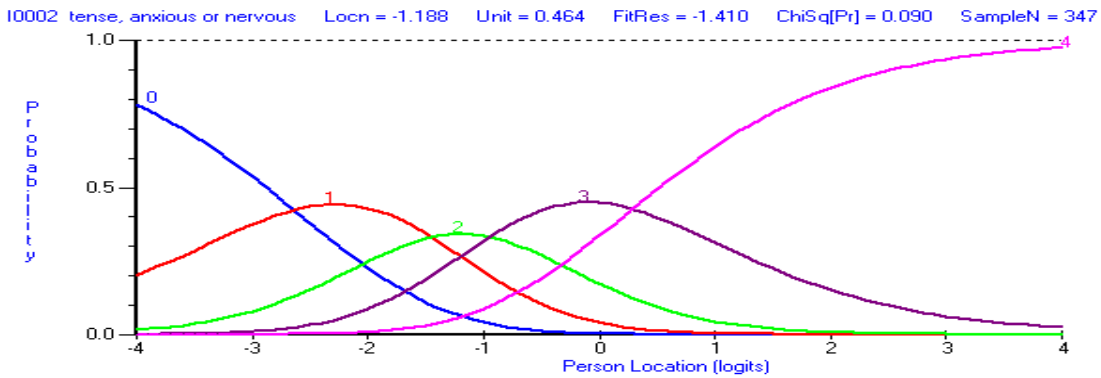
Table A11. Findings of Principal Components Analysis of CORE-OM data in study sample [N750b]. Oblique rotation – component correlation matrix

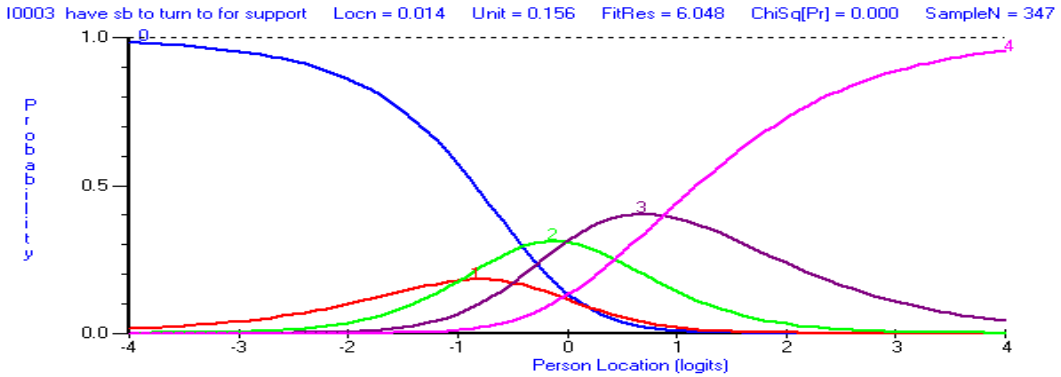
Component	1	2	3	4	5	6	7
1	**1.00**	**0.60**	**0.58**	0.37	**0.57**	0.20	0.04
2	**0.60**	**1.00**	**0.52**	0.32	**0.54**	0.17	0.17
3	**0.58**	**0.52**	**1.00**	0.31	**0.45**	0.07	0.06
4	0.37	0.32	0.31	**1.00**	0.38	0.13	0.15
5	**0.57**	**0.54**	**0.45**	0.38	**1.00**	0.17	0.18
6	0.20	0.17	0.07	0.13	0.17	**1.00**	0.03
7	0.04	0.17	0.06	0.15	0.18	0.03	**1.00**

Rotation method Promax with Kaiser normalisation. Correlation coefficients ≥ │0.40│ are shown in bold; correlations of the 3 components that were found to be non-significant by Horn's analysis have been shaded in grey.

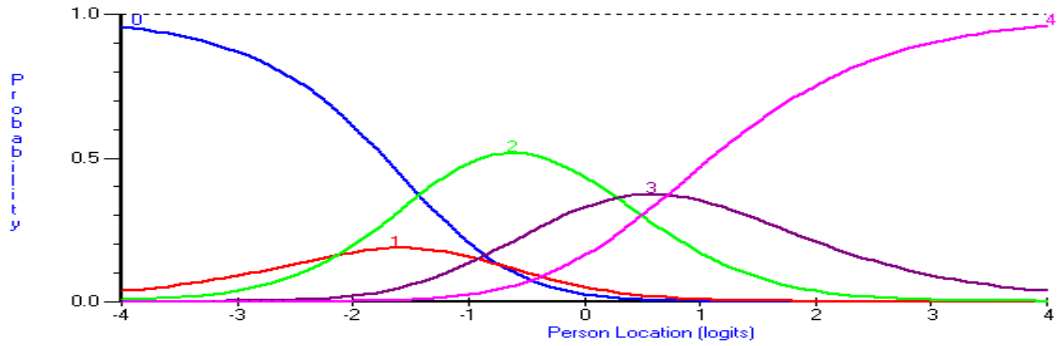
Appendix 6. Rasch category probability curves of the 34 CORE-OM items before rescoring - [N400a] dataset



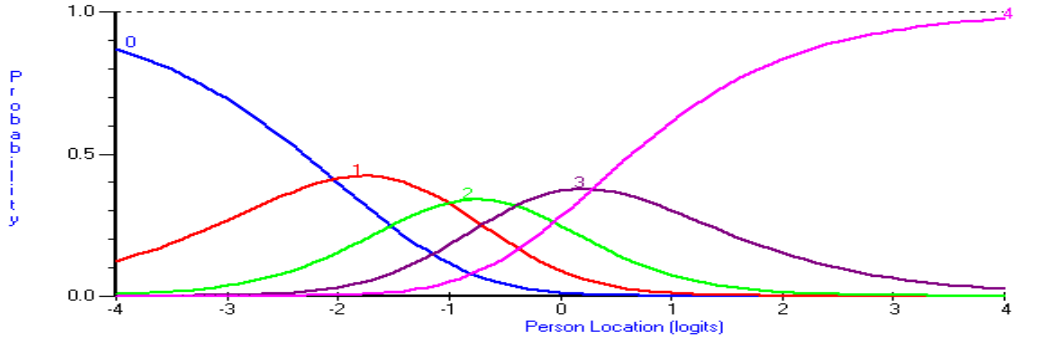




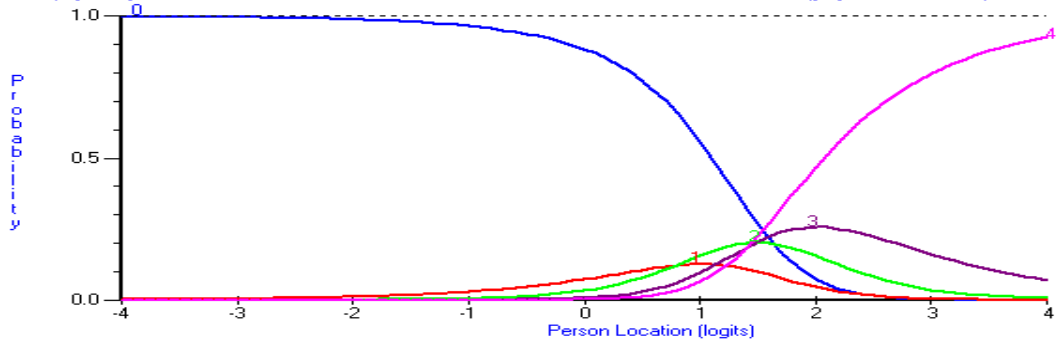
I0004 felt ok about myself Locn = -0.464 Unit = 0.337 FitRes = -1.308 ChiSq[Pr] = 0.019 SampleN = 347



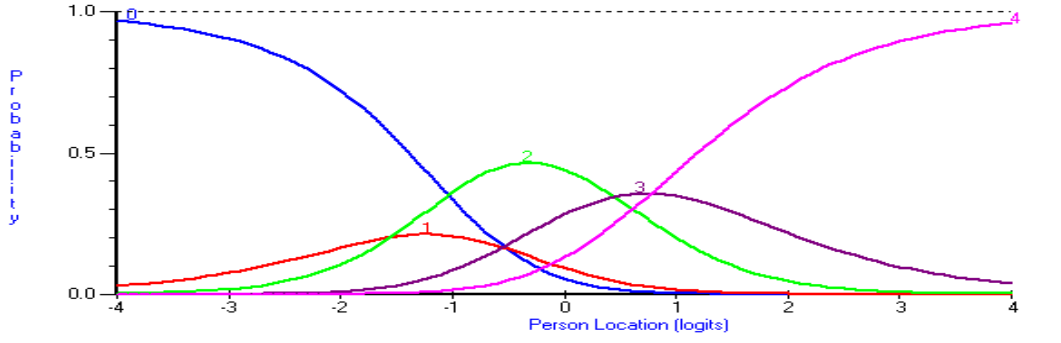
I0005 totally lacking in energy/enth Locn = -0.790 Unit = 0.381 FitRes = 0.392 ChiSq[Pr] = 0.214 SampleN = 347



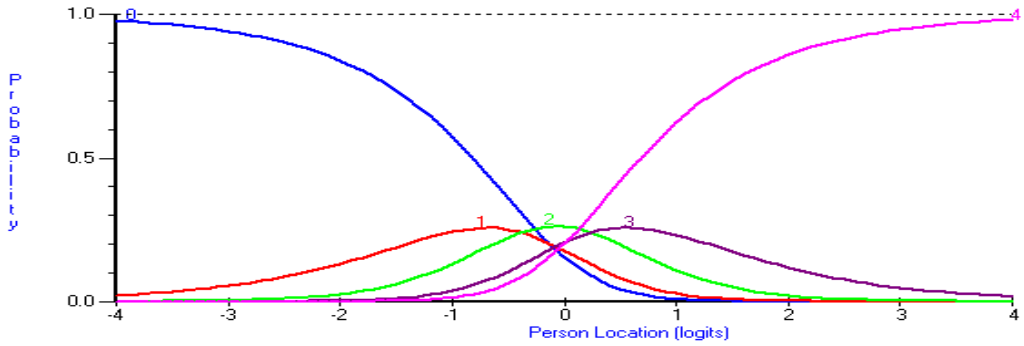
I0006 physically violent to others Locn = 1.544 Unit = -0.127 FitRes = 0.077 ChiSq[Pr] = 0.491 SampleN = 347



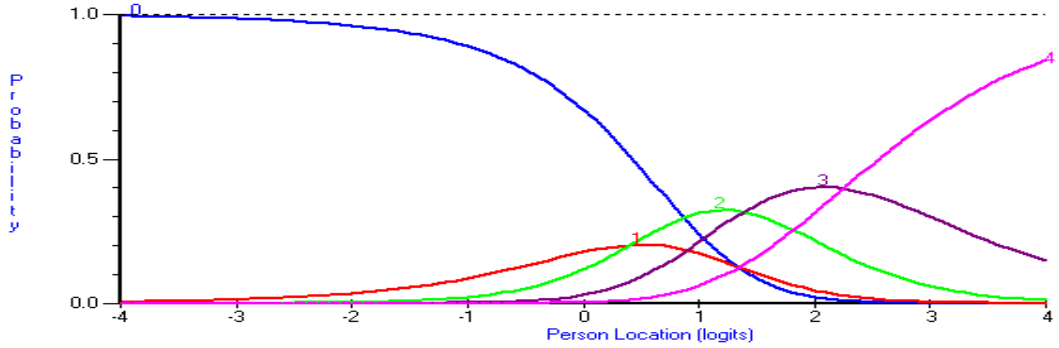
I0007 able to cope when things go wr Locn = -0.212 Unit = 0.294 FitRes = 0.027 ChiSq[Pr] = 0.662 SampleN = 347



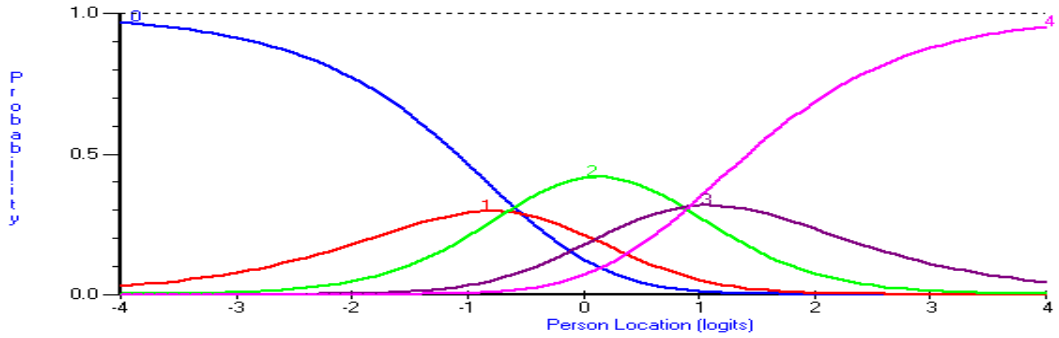
I0008 troubled by aches, pains, phys Locn = -0.067 Unit = 0.055 FitRes = 7.976 ChiSq[Pr] = 0.000 SampleN = 347



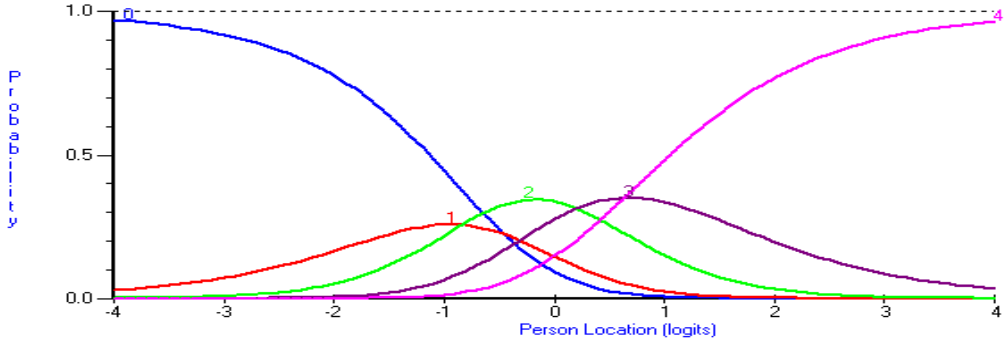
I0009 thought of hurting myself Locn = 1.344 Unit = 0.186 FitRes = -1.037 ChiSq[Pr] = 0.209 SampleN = 347



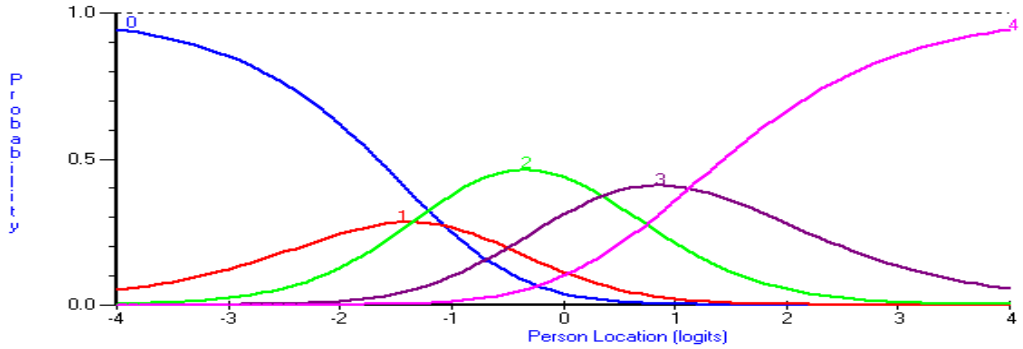
I0010 talking to people felt too muc Locn = 0.148 Unit = 0.294 FitRes = 0.113 ChiSq[Pr] = 0.077 SampleN = 347



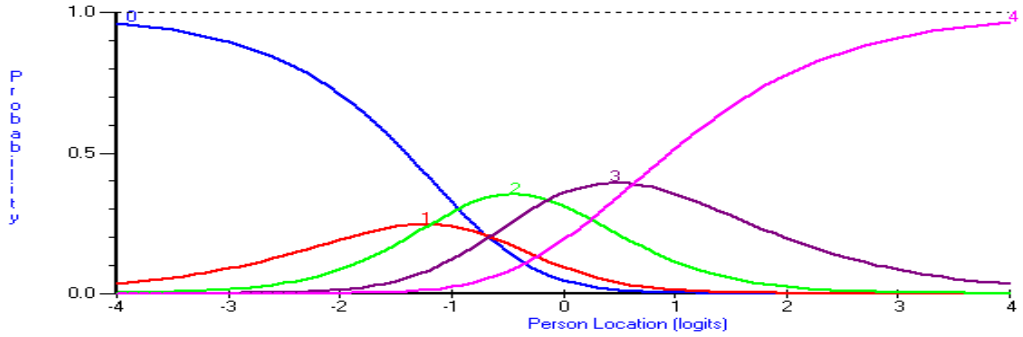
I0011 tension/anxiety prevented me f Locn = -0.111 Unit = 0.216 FitRes = -0.383 ChiSq[Pr] = 0.758 SampleN = 347



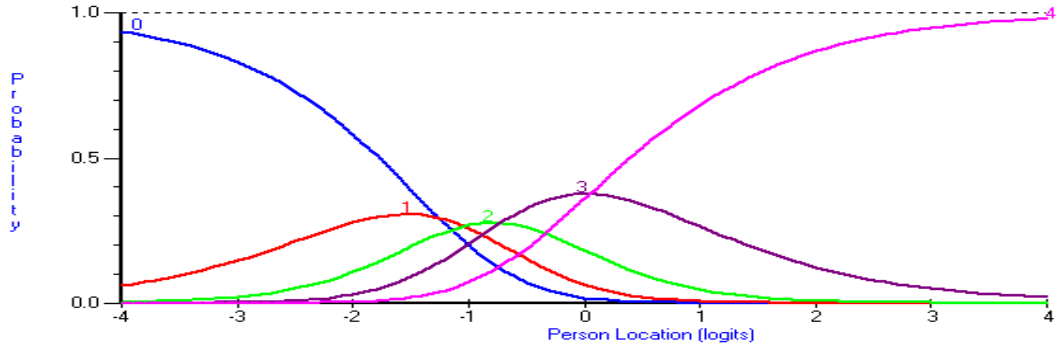
I0012 happy with the things I've don Locn = -0.232 Unit = 0.415 FitRes = -0.650 ChiSq[Pr] = 0.551 SampleN = 347



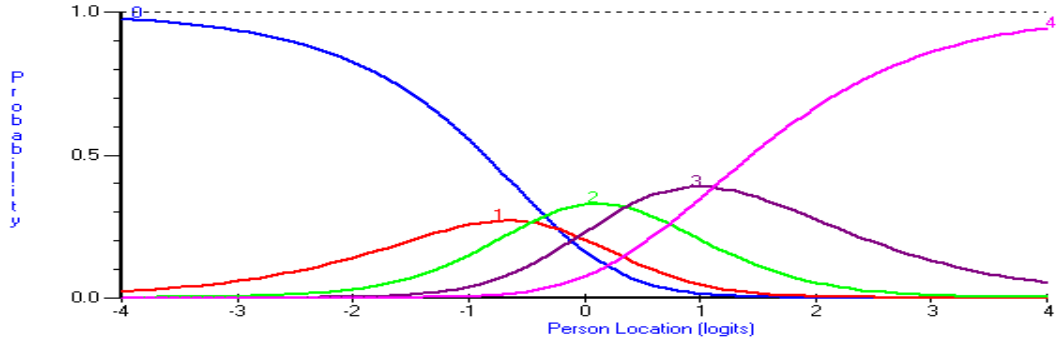
I0013 disturbed by unwanted thoughts Locn = -0.347 Unit = 0.250 FitRes = 1.399 ChiSq[Pr] = 0.601 SampleN = 347



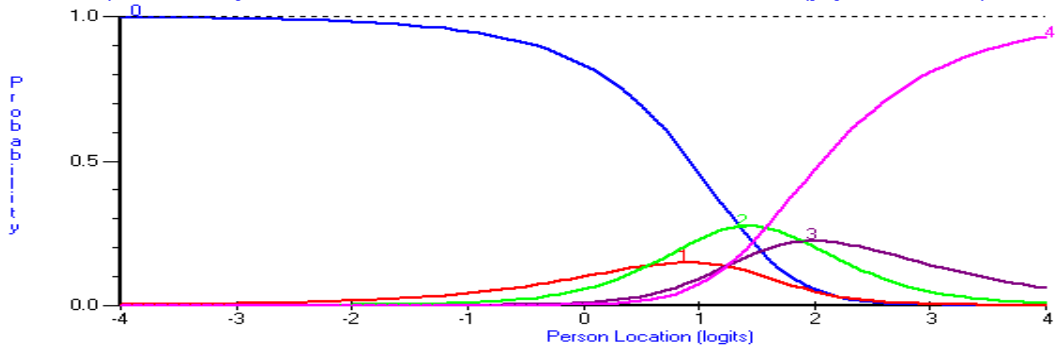
I0014 felt like crying Locn = -0.745 Unit = 0.213 FitRes = -1.770 ChiSq[Pr] = 0.009 SampleN = 347



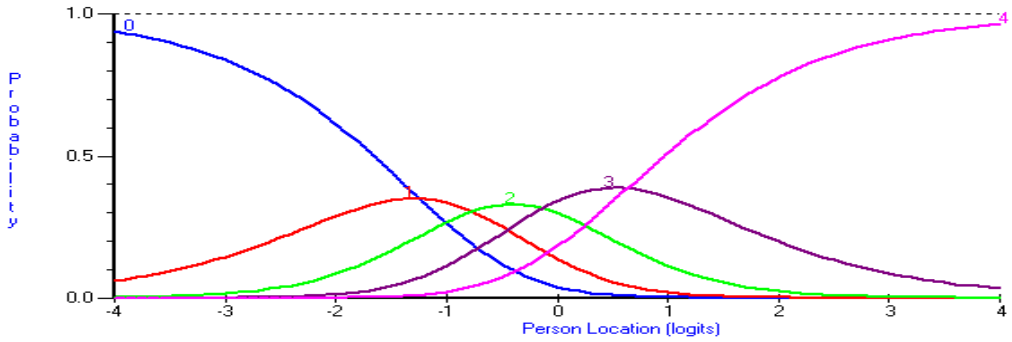
I0015 felt panic or terror Locn = 0.196 Unit = 0.242 FitRes = 0.435 ChiSq[Pr] = 0.986 SampleN = 347



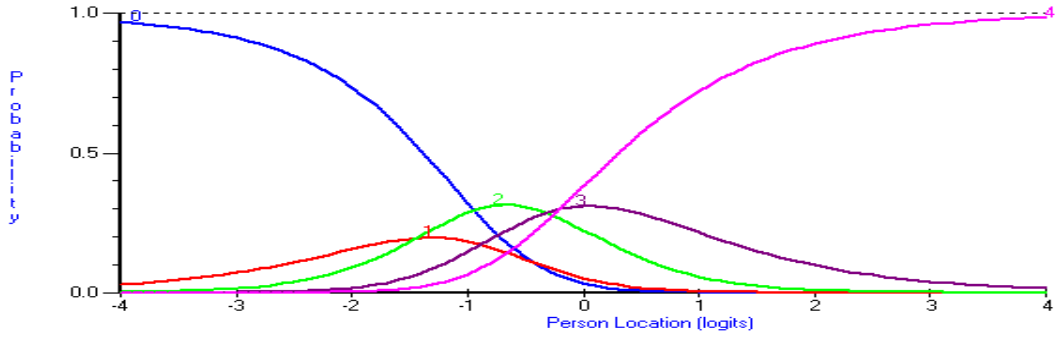
I0016 made plans to end my life Locn = 1.464 Unit = -0.066 FitRes = -0.758 ChiSq[Pr] = 0.685 SampleN = 347



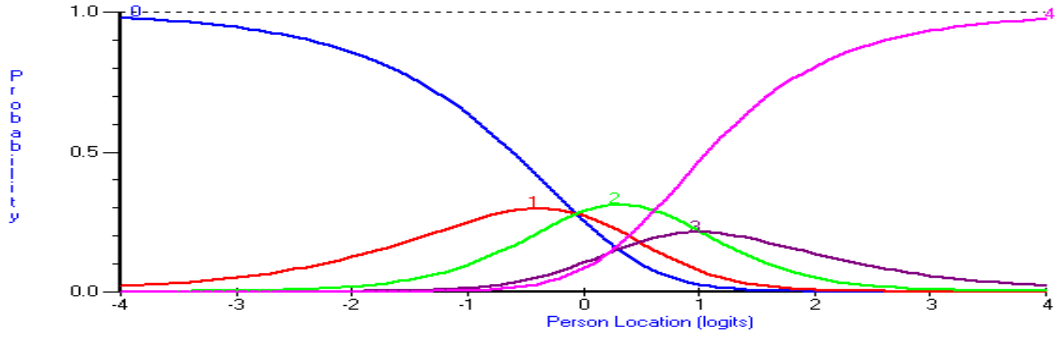
I0017 overwhelmed by my problems Locn = -0.379 Unit = 0.314 FitRes = -2.786 ChiSq[Pr] = 0.000 SampleN = 347



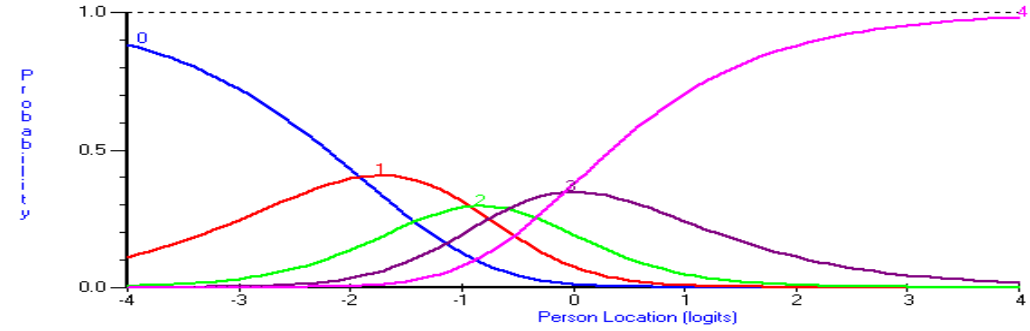
I0018 difficulty getting to sleep Locn = -0.606 Unit = 0.092 FitRes = 2.937 ChiSq[Pr] = 0.026 SampleN = 347



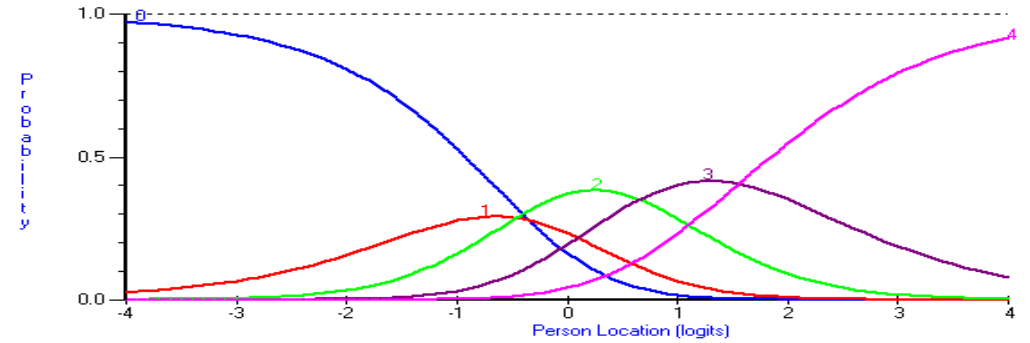
I0019 felt warmth/affection for sb Locn = 0.282 Unit = 0.096 FitRes = 7.368 ChiSq[Pr] = 0.000 SampleN = 347



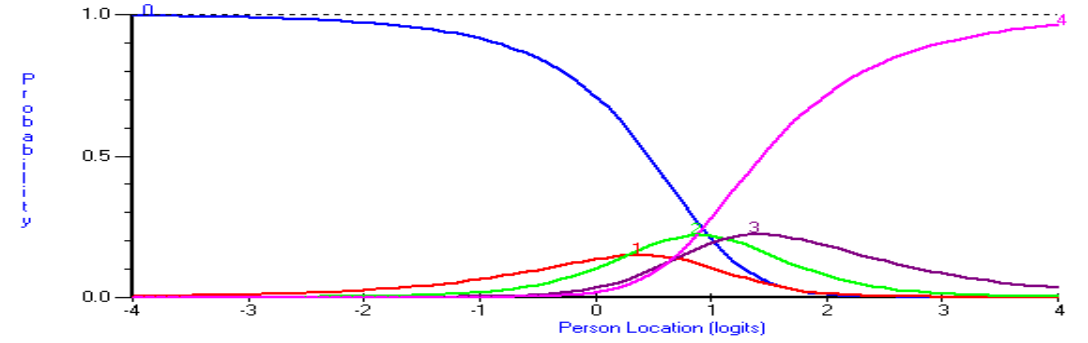
I0020 problems impossible to put to Locn = -0.877 Unit = 0.293 FitRes = -1.100 ChiSq[Pr] = 0.265 SampleN = 347



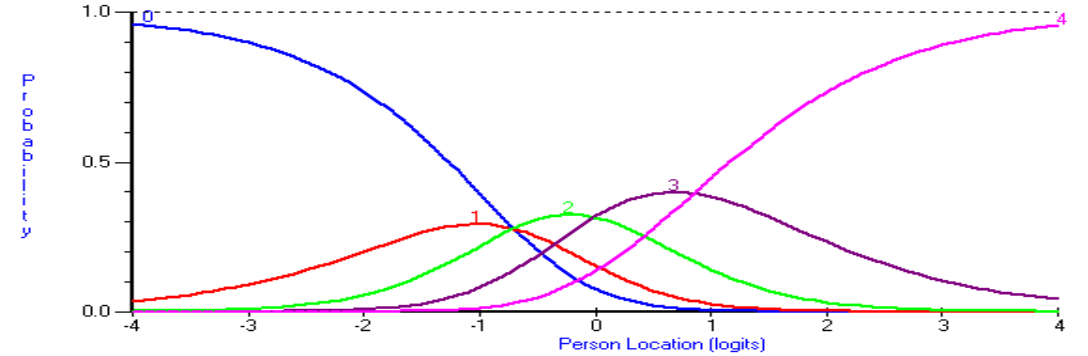
I0021 able to do most things needed Locn = 0.349 Unit = 0.342 FitRes = -0.413 ChiSq[Pr] = 0.301 SampleN = 347



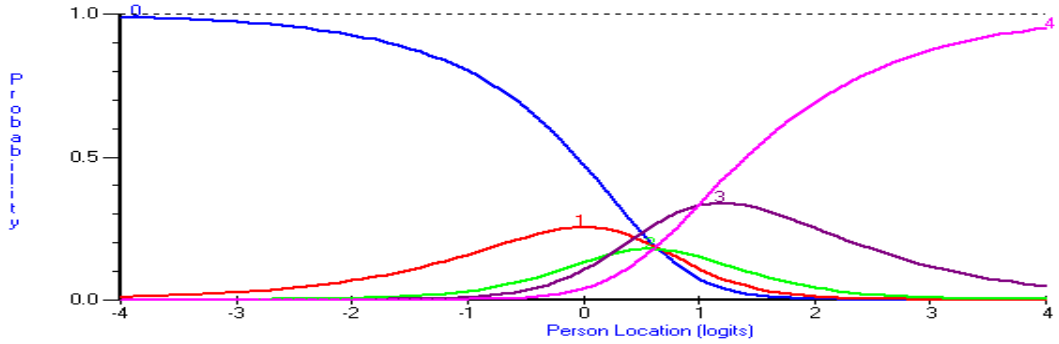
I0022 threatened/intimidated sb Locn = 0.927 Unit = -0.113 FitRes = 2.670 ChiSq[Pr] = 0.000 SampleN = 347



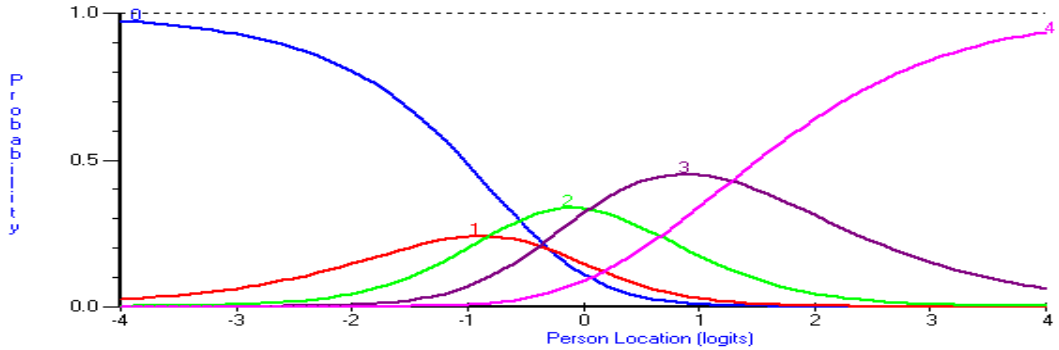
I0023 despairing or hopeless Locn = -0.141 Unit = 0.269 FitRes = -4.447 ChiSq[Pr] = 0.000 SampleN = 347



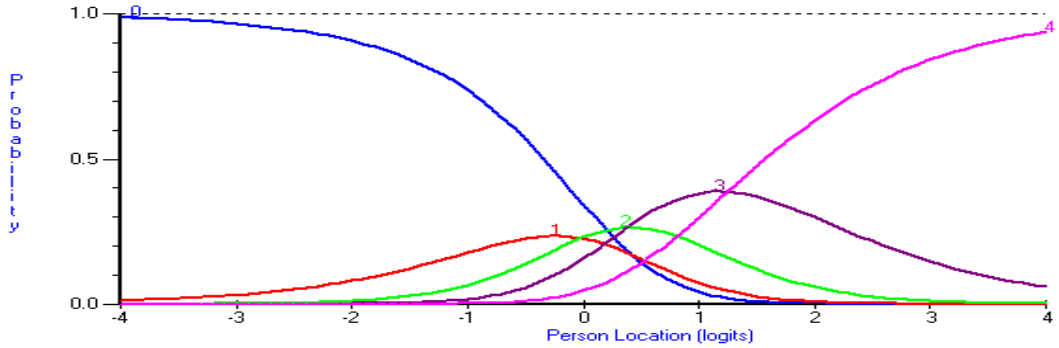
I0024 thought better if dead Locn = 0.628 Unit = 0.033 FitRes = -2.677 ChiSq[Pr] = 0.033 SampleN = 347



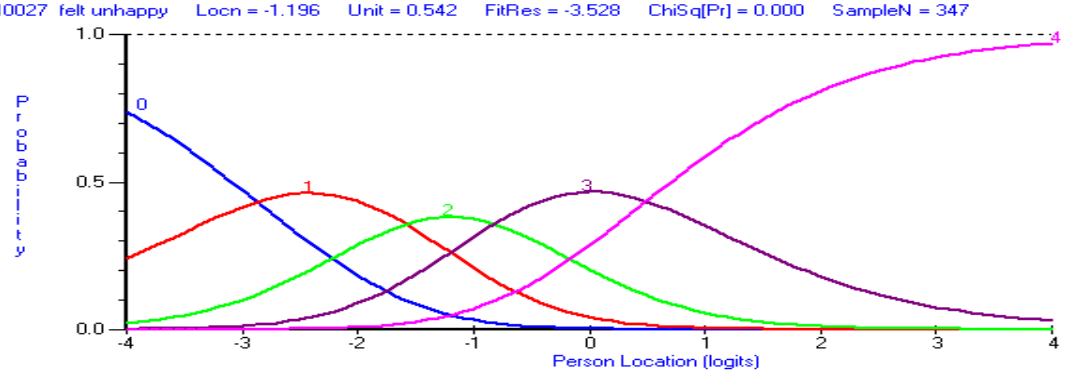
I0025 felt criticised by other peopl Locn = 0.056 Unit = 0.279 FitRes = 0.703 ChiSq[Pr] = 0.136 SampleN = 347



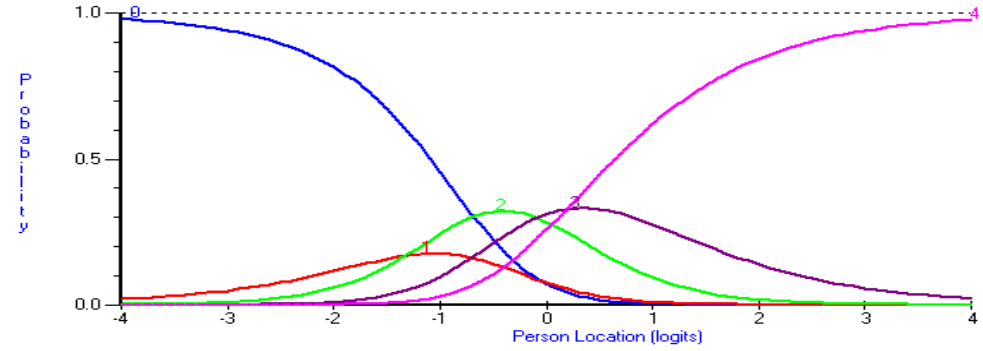
I0026 thought I have no friends Locn = 0.509 Unit = 0.143 FitRes = 0.218 ChiSq[Pr] = 0.755 SampleN = 347



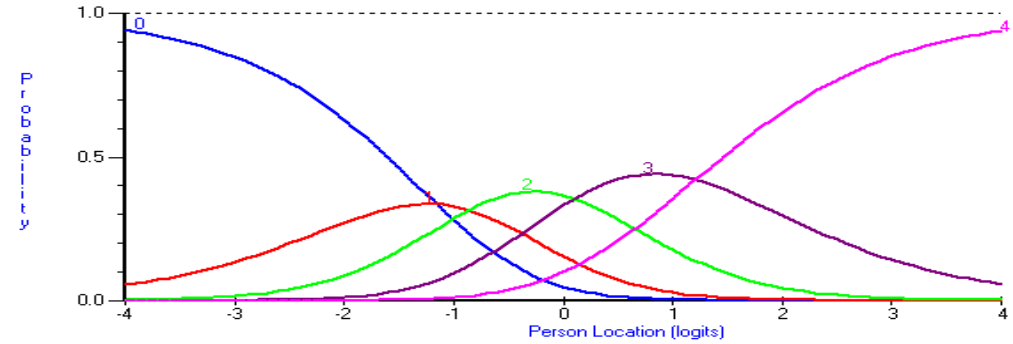
I0027 felt unhappy Locn = -1.196 Unit = 0.542 FitRes = -3.528 ChiSq[Pr] = 0.000 SampleN = 347



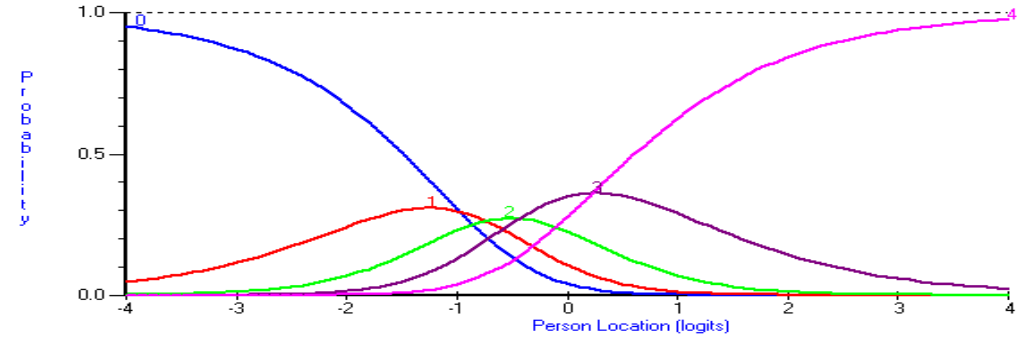
I0028 unwanted images/memories distr Locn = -0.316 Unit = 0.098 FitRes = 1.782 ChiSq[Pr] = 0.015 SampleN = 347



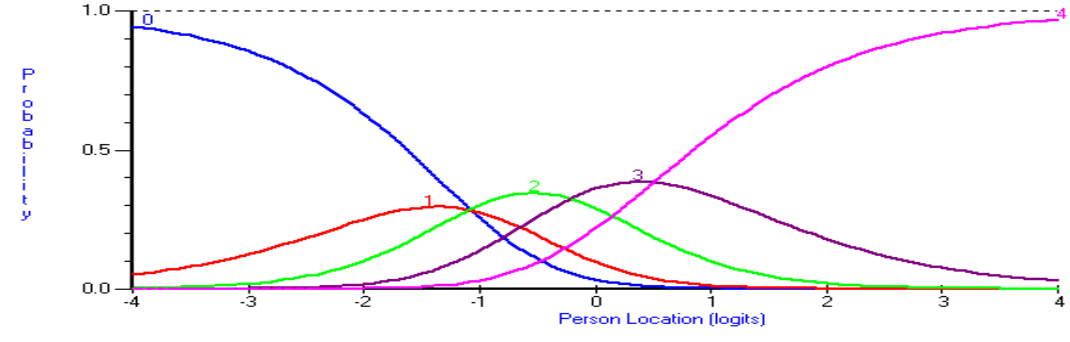
I0029 irritable when with other peop Locn = -0.177 Unit = 0.403 FitRes = 0.879 ChiSq[Pr] = 0.931 SampleN = 347



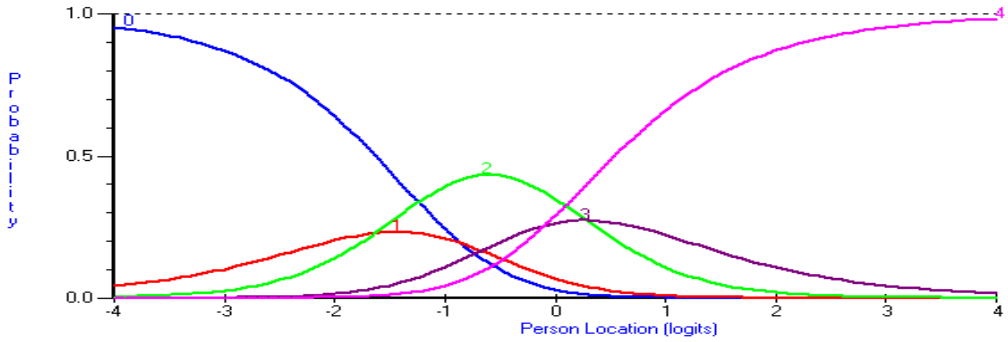
I0030 thought I am to blame for my p Locn = -0.478 Unit = 0.196 FitRes = 3.669 ChiSq[Pr] = 0.017 SampleN = 347



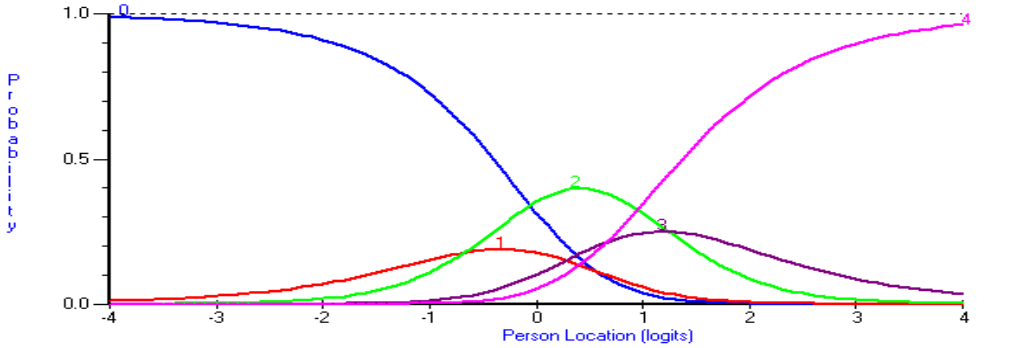
I0031 optimistic about future Locn = -0.468 Unit = 0.281 FitRes = 4.857 ChiSq[Pr] = 0.002 SampleN = 347



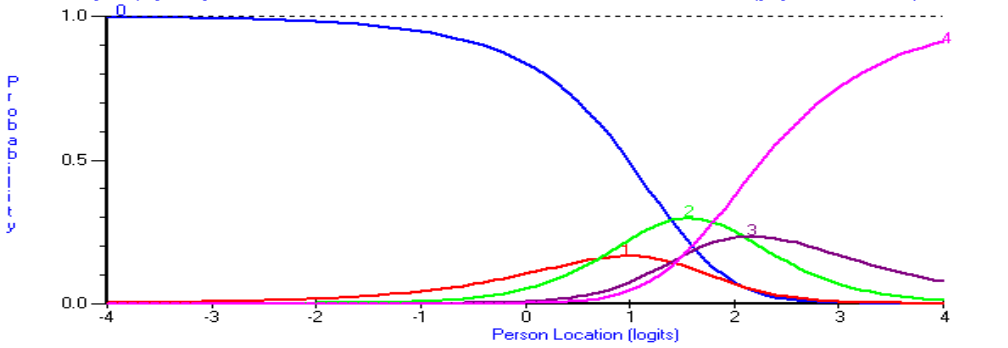
I0032 achieved things I wanted to Locn = -0.573 Unit = 0.210 FitRes = -0.049 ChiSq[Pr] = 0.858 SampleN = 347



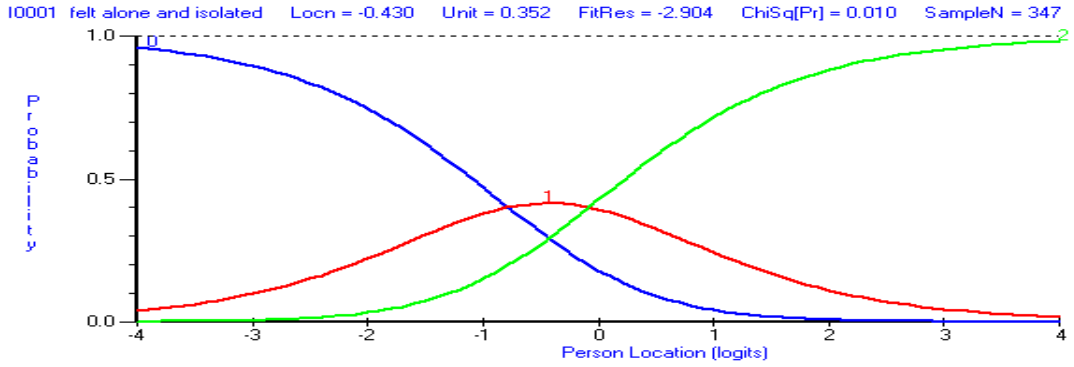
I0033 felt humiliated/shamed by othe Locn = 0.448 Unit = 0.109 FitRes = -0.823 ChiSq[Pr] = 0.413 SampleN = 347

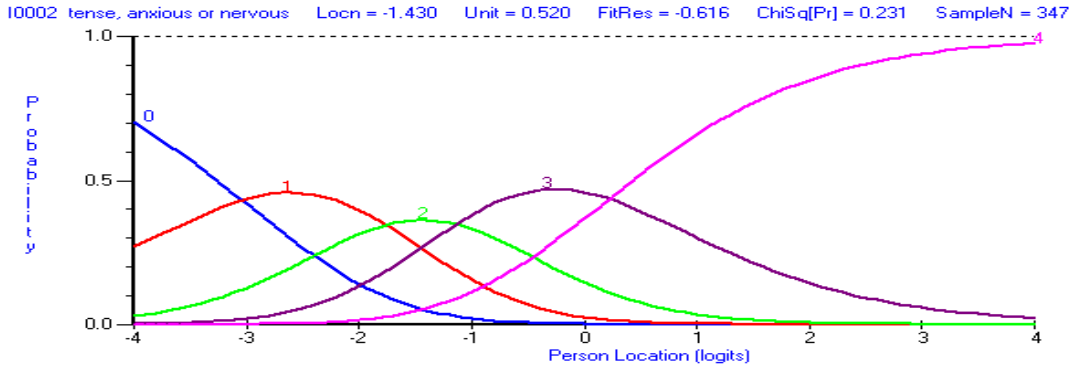


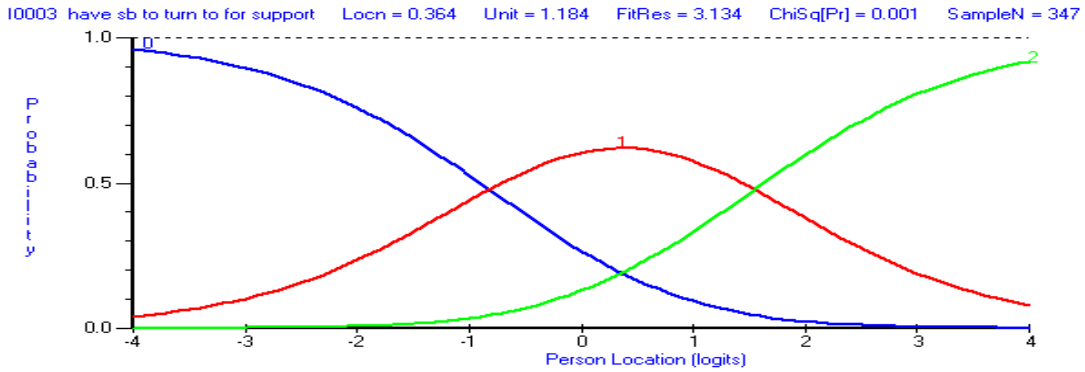
I0034 hurt myself physically/taken r Locn = 1.598 Unit = -0.018 FitRes = -1.204 ChiSq[Pr] = 0.576 SampleN = 347



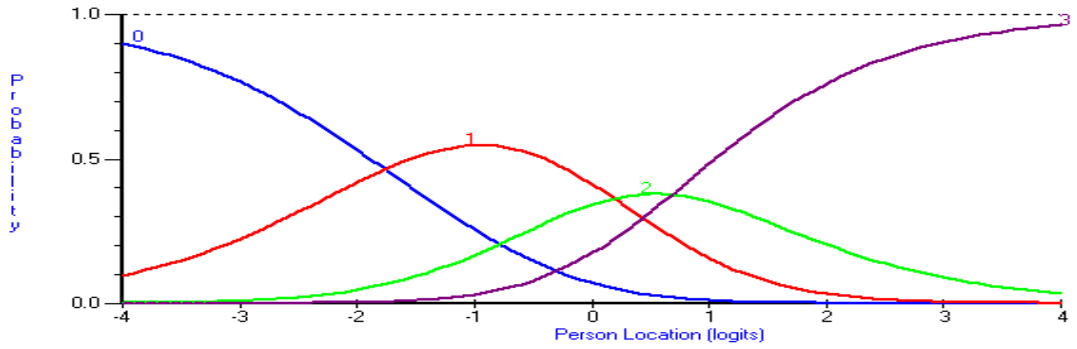
Appendix 7. Rasch category probability curves for the 34 CORE-OM items after rescoring – Rasch analysis on [N400a]



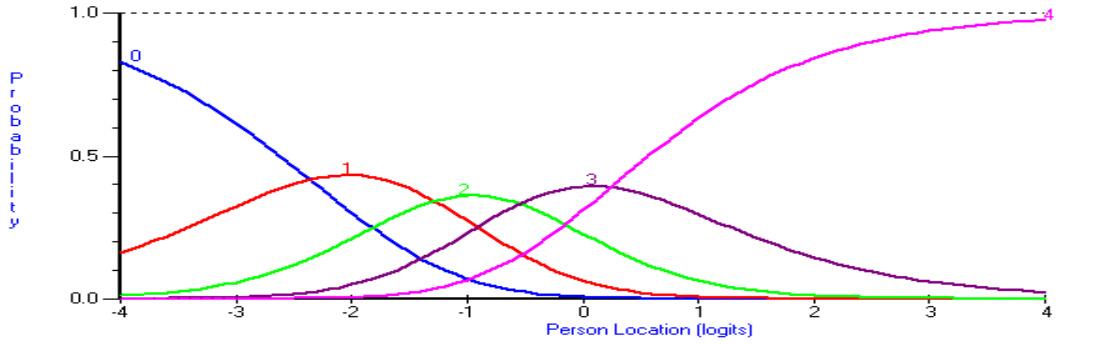




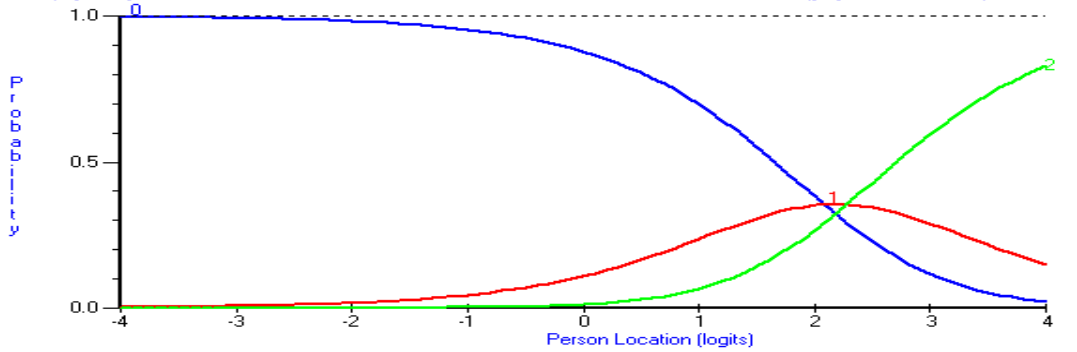
I0004 felt ok about myself Locn = -0.296 Unit = 0.611 FitRes = -1.278 ChiSq[Pr] = 0.051 SampleN = 347



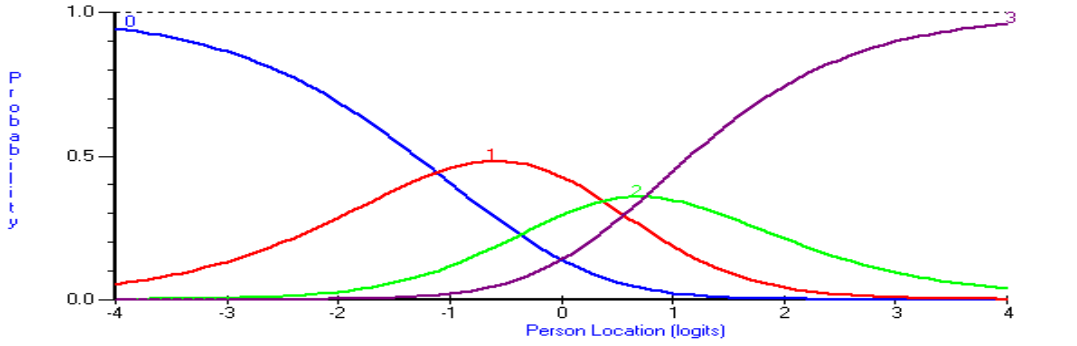
I0005 totally lacking in energy/enth Locn = -0.986 Unit = 0.427 FitRes = 1.630 ChiSq[Pr] = 0.083 SampleN = 347

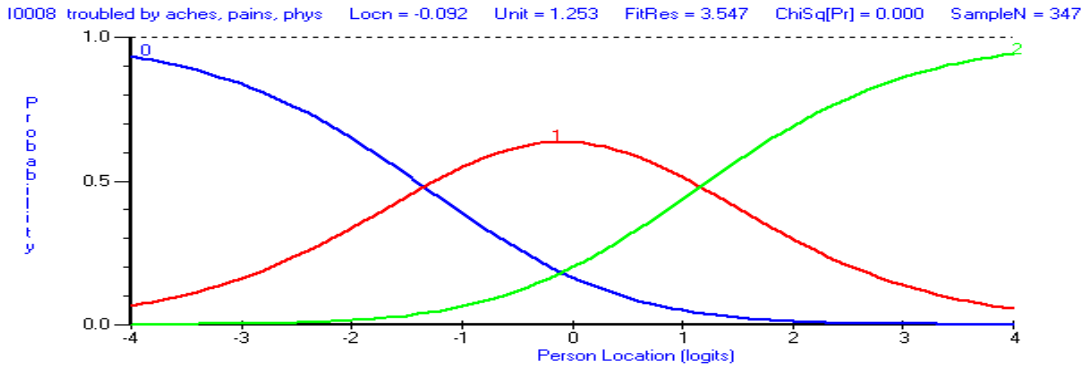


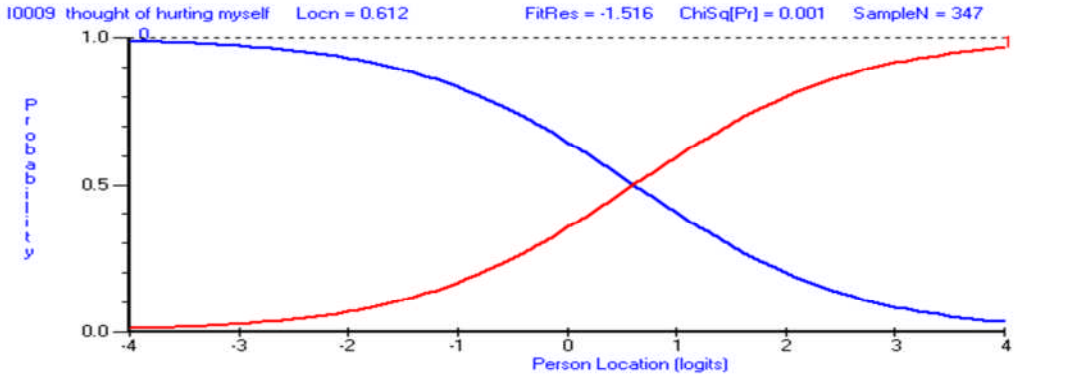
I0006 physically violent to others Locn = 2.182 Unit = 0.097 FitRes = -0.394 ChiSq[Pr] = 0.018 SampleN = 347

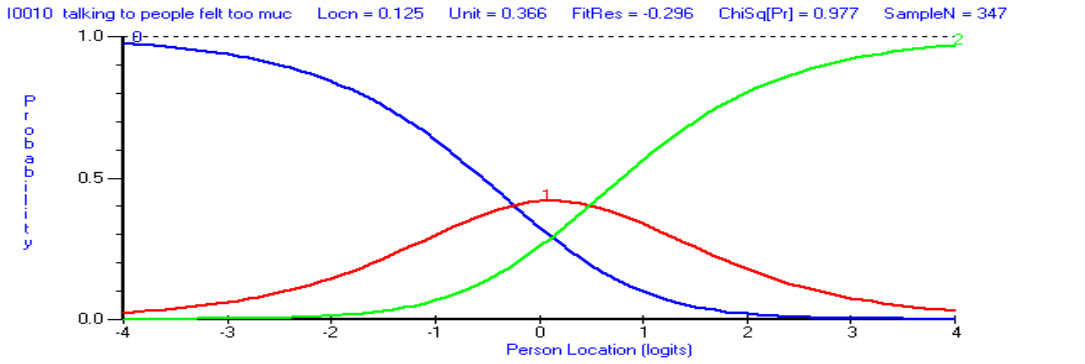


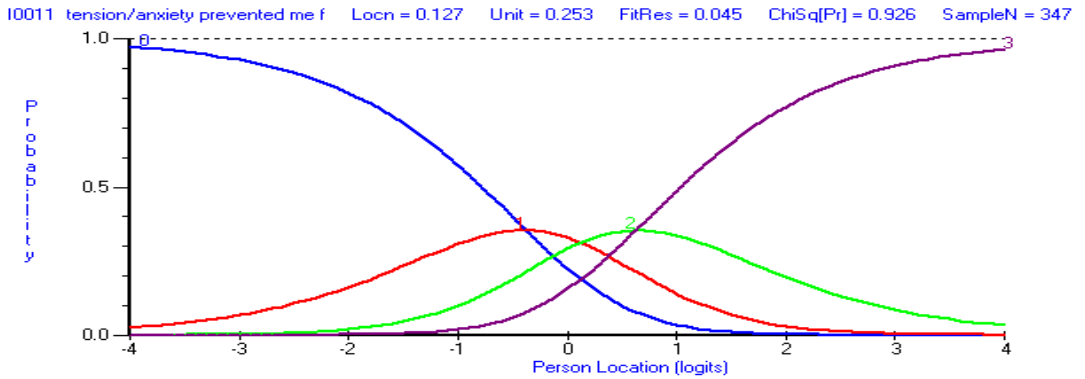
I0007 able to cope when things go wr Locn = 0.003 Unit = 0.467 FitRes = 0.619 ChiSq[Pr] = 0.451 SampleN = 347



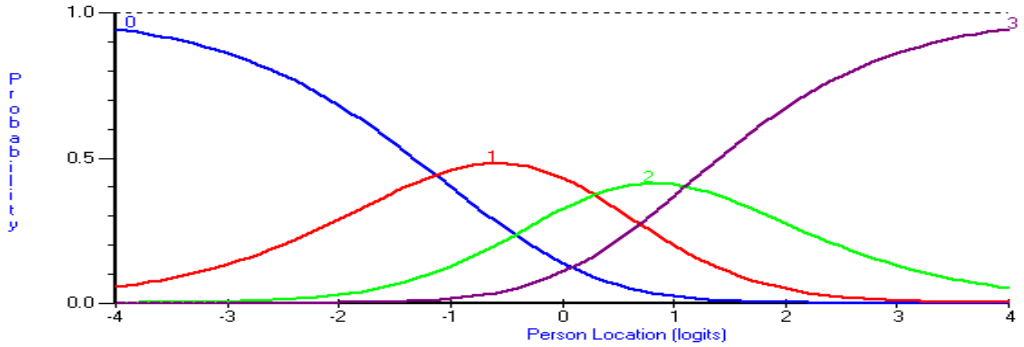

I0008 troubled by aches, pains, phys Locn = -0.092 Unit = 1.253 FitRes = 3.547 ChiSq[Pr] = 0.000 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)


I0009 thought of hurting myself Locn = 0.612 FitRes = -1.516 ChiSq[Pr] = 0.001 SampleN = 347
Probability
1.0
0.5
0.0
0
1
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)

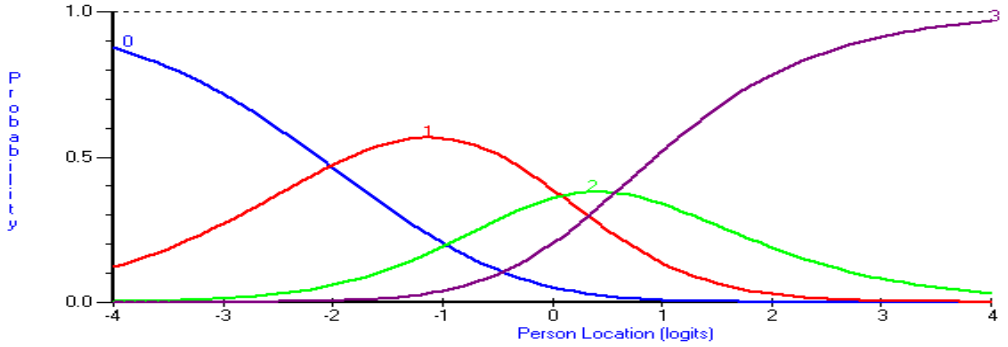

I0010 talking to people felt too muc Locn = 0.125 Unit = 0.366 FitRes = -0.296 ChiSq[Pr] = 0.977 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)


I0011 tension/anxiety prevented me f Locn = 0.127 Unit = 0.253 FitRes = 0.045 ChiSq[Pr] = 0.926 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
3
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)

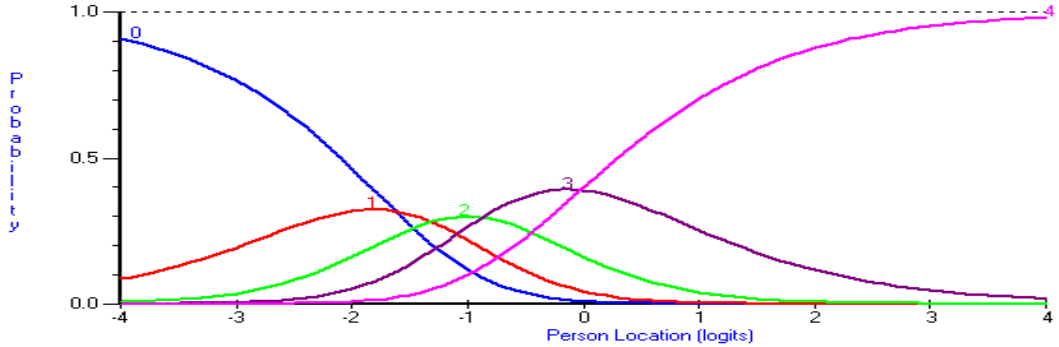
I0012 happy with the things I've don Locn = 0.086 Unit = 0.561 FitRes = 0.177 ChiSq[Pr] = 0.888 SampleN = 347



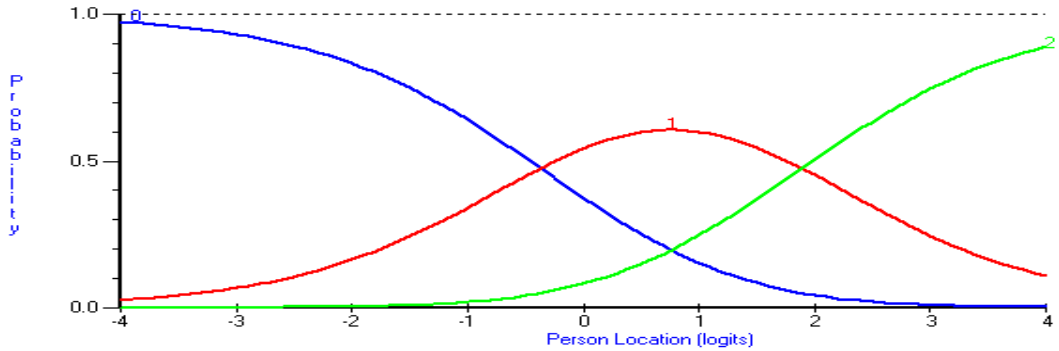
I0013 disturbed by unwanted thoughts Locn = -0.452 Unit = 0.648 FitRes = 1.416 ChiSq[Pr] = 0.049 SampleN = 347

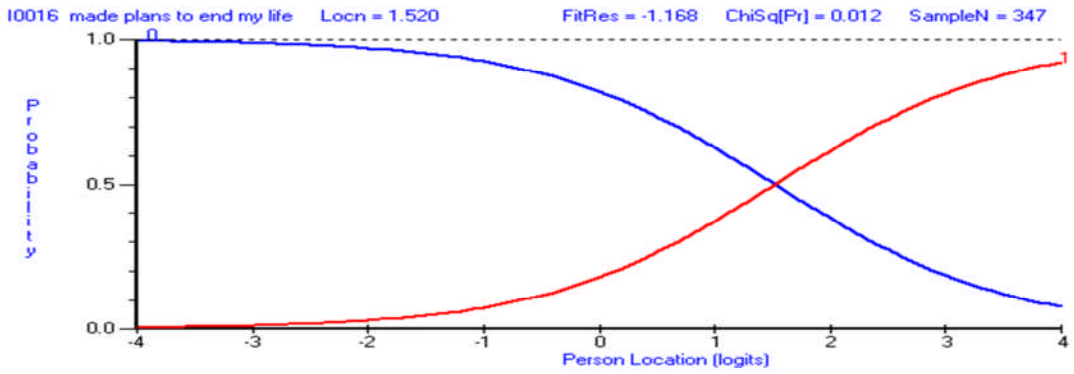


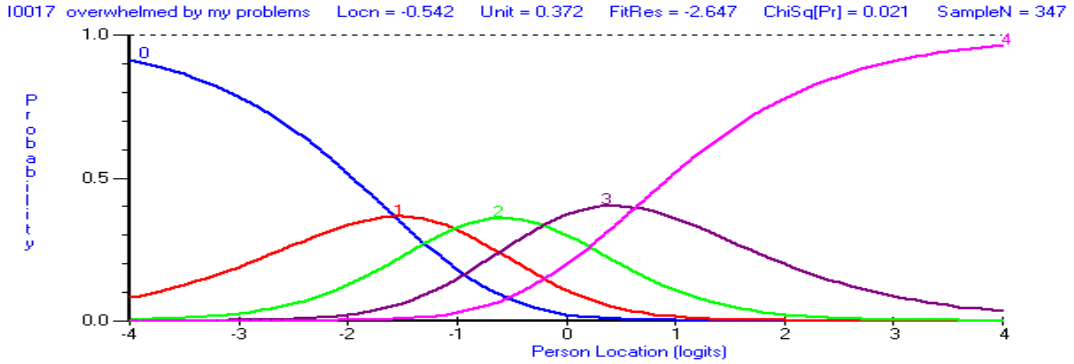
I0014 felt like crying Locn = -0.958 Unit = 0.262 FitRes = -1.024 ChiSq[Pr] = 0.857 SampleN = 347

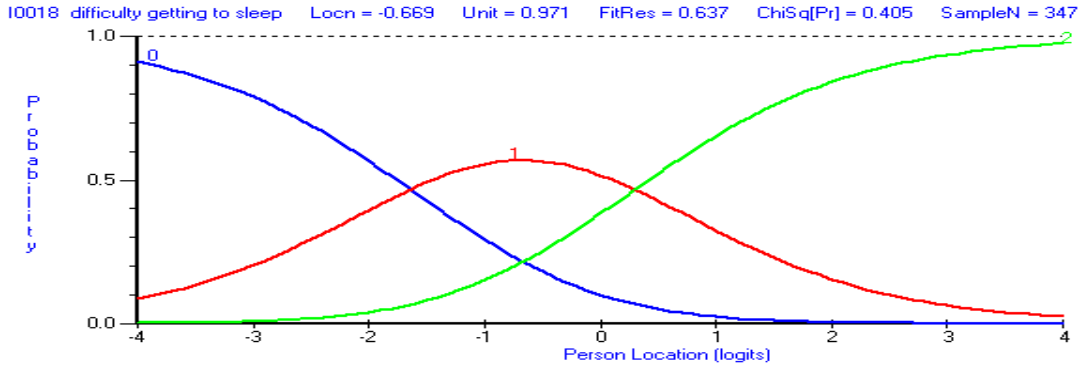


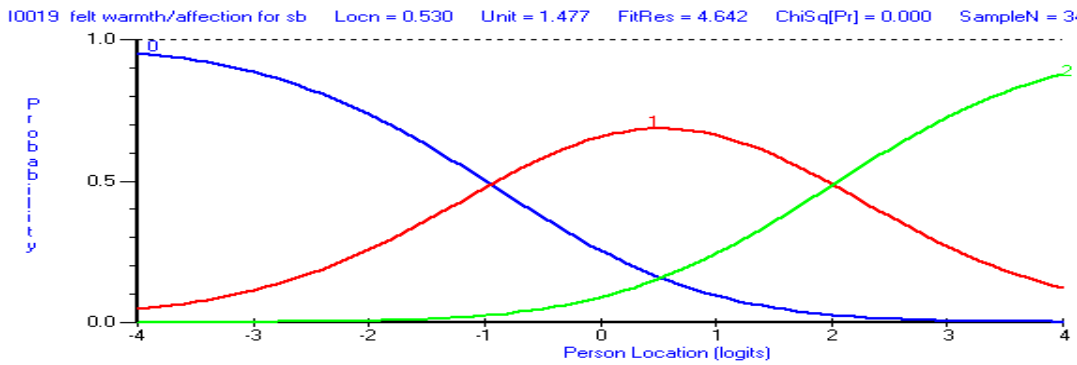
I0015 felt panic or terror Locn = 0.760 Unit = 1.128 FitRes = -0.083 ChiSq[Pr] = 0.653 SampleN = 347



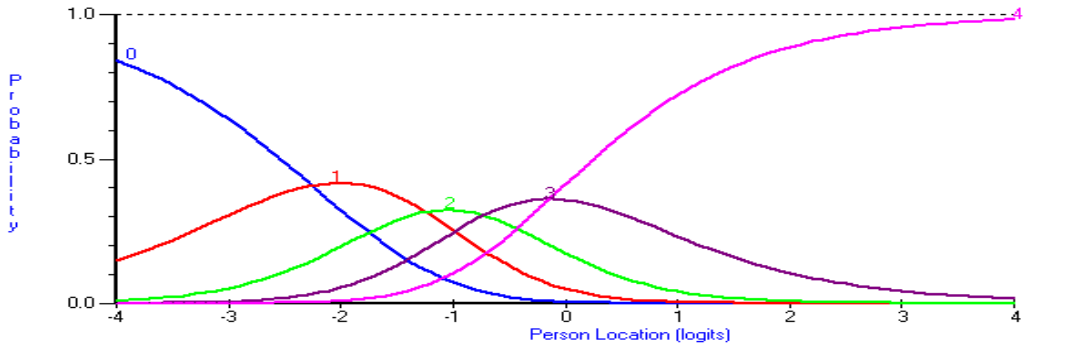

I0016 made plans to end my life Locn = 1.520 FitRes = -1.168 ChiSq[Pr] = 0.012 SampleN = 347
Probability
1.0
0.5
0.0
0
1
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)


I0017 overwhelmed by my problems Locn = -0.542 Unit = 0.372 FitRes = -2.647 ChiSq[Pr] = 0.021 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
3
4
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)

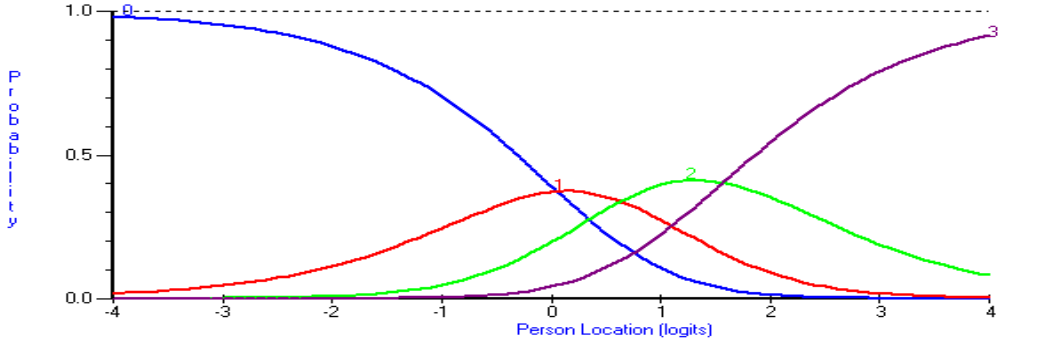

I0018 difficulty getting to sleep Locn = -0.669 Unit = 0.971 FitRes = 0.637 ChiSq[Pr] = 0.405 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)


I0019 felt warmth/affection for sb Locn = 0.530 Unit = 1.477 FitRes = 4.642 ChiSq[Pr] = 0.000 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)

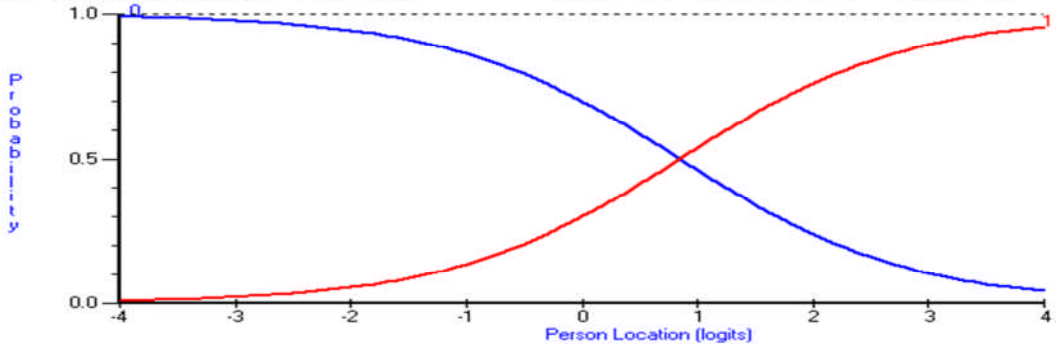
I0020 problems impossible to put to Locn = -1.090 Unit = 0.345 FitRes = 0.276 ChiSq[Pr] = 0.939 SampleN = 347



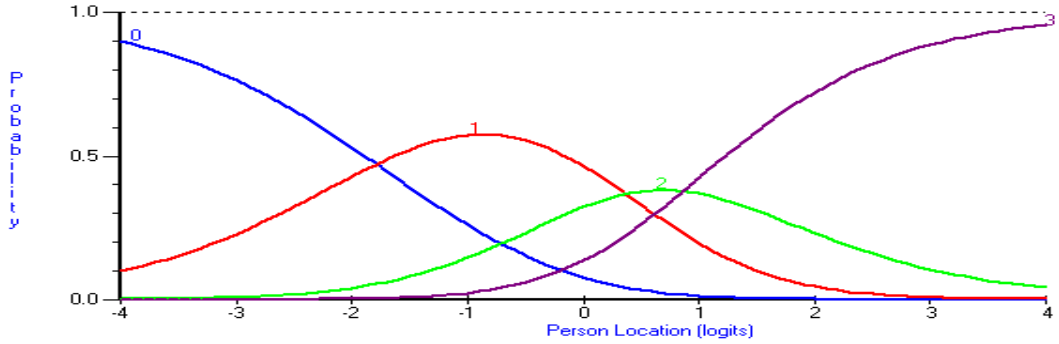
I0021 able to do most things needed Locn = 0.753 Unit = 0.378 FitRes = -0.591 ChiSq[Pr] = 0.071 SampleN = 347



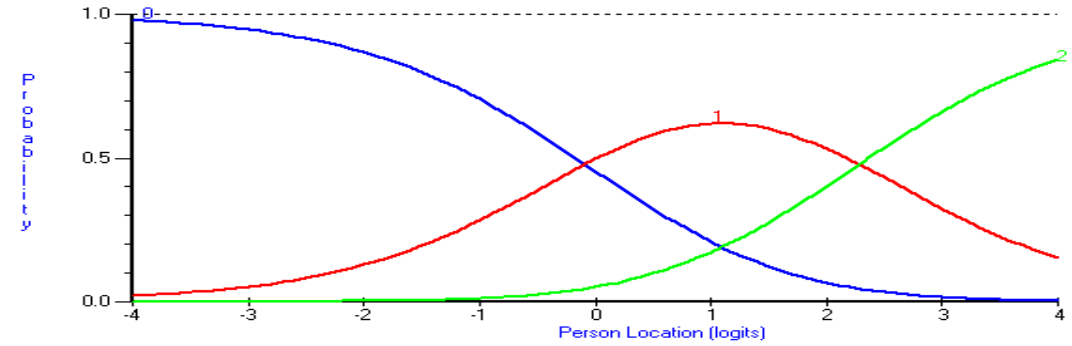
I0022 threatened/intimidated sb Locn = 0.847 FitRes = 1.259 ChiSq[Pr] = 0.145 SampleN = 347



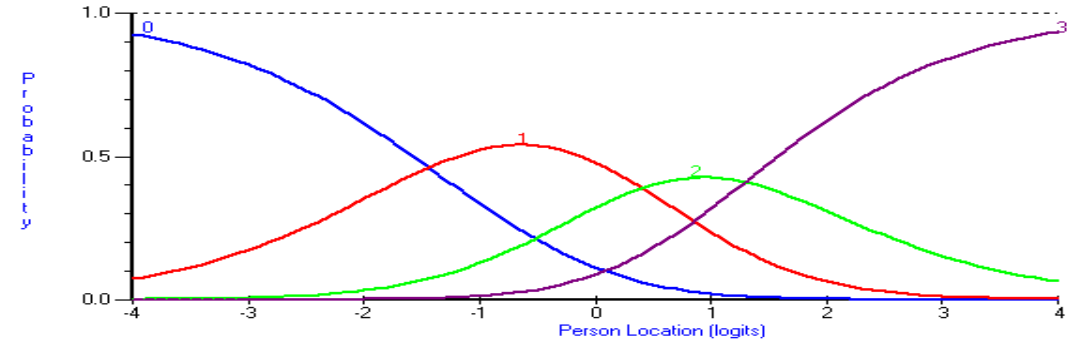
I0023 despairing or hopeless Locn = -0.186 Unit = 0.663 FitRes = -4.230 ChiSq[Pr] = 0.000 SampleN = 347



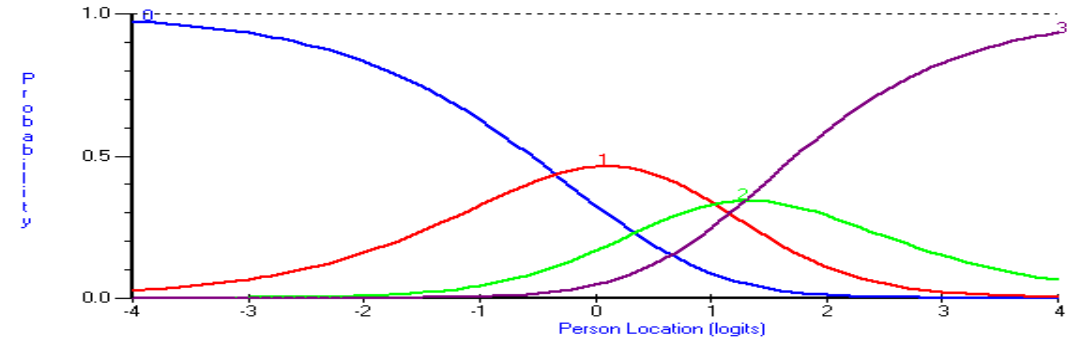
I0024 thought better if dead Locn = 1.098 Unit = 1.186 FitRes = -1.926 ChiSq[Pr] = 0.001 SampleN = 347



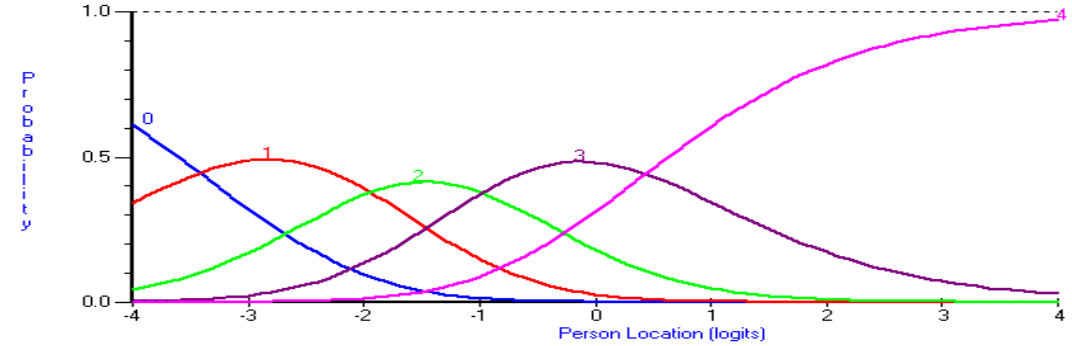
I0025 felt criticised by other peopl Locn = 0.087 Unit = 0.685 FitRes = -0.445 ChiSq[Pr] = 0.384 SampleN = 347



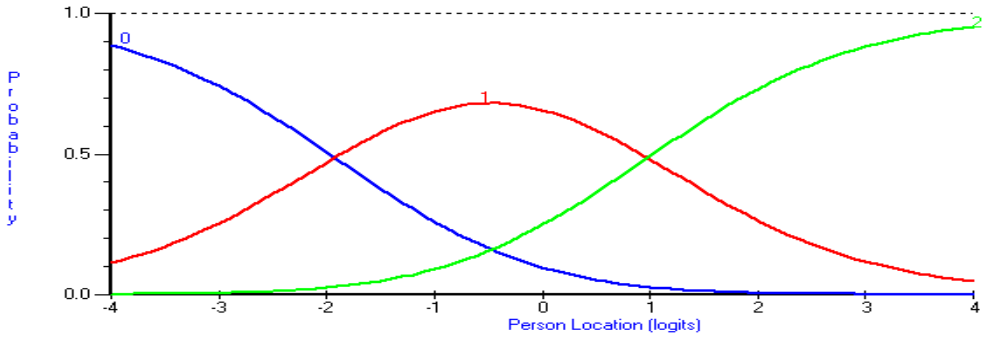
I0026 thought I have no friends Locn = 0.660 Unit = 0.409 FitRes = -0.238 ChiSq[Pr] = 0.351 SampleN = 347



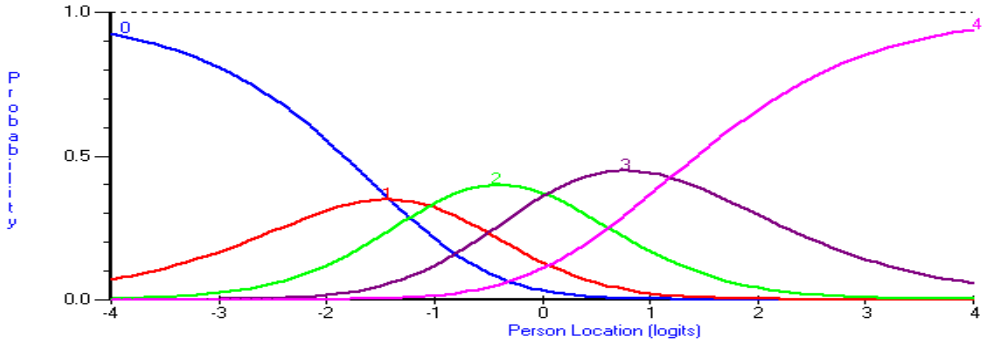
I0027 felt unhappy Locn = -1.476 Unit = 0.627 FitRes = -4.086 ChiSq[Pr] = 0.000 SampleN = 347



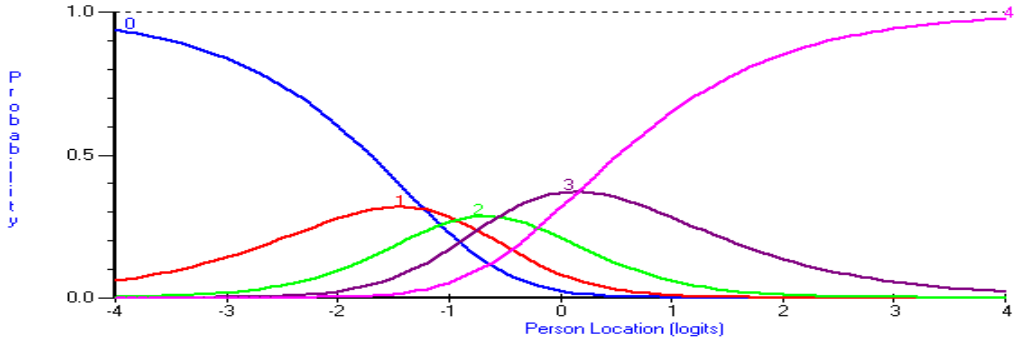
I0028 unwanted images/memories distr Locn = -0.472 Unit = 1.451 FitRes = -0.582 ChiSq[Pr] = 0.020 SampleN = 347



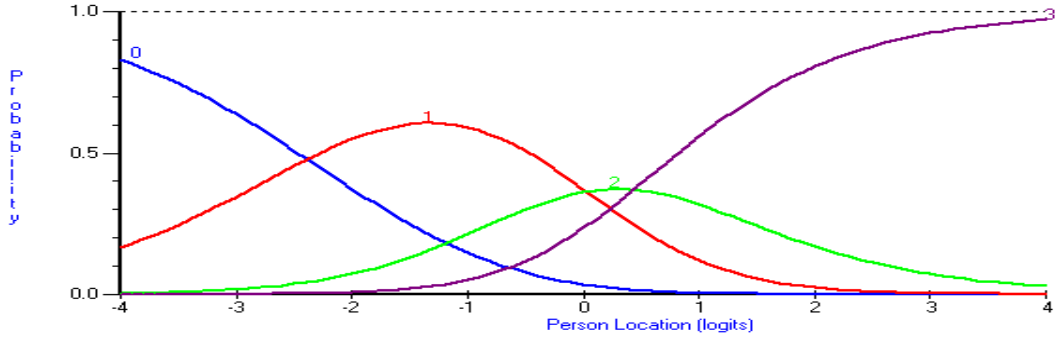
I0029 irritable when with other peop Locn = -0.307 Unit = 0.445 FitRes = 2.491 ChiSq[Pr] = 0.082 SampleN = 347

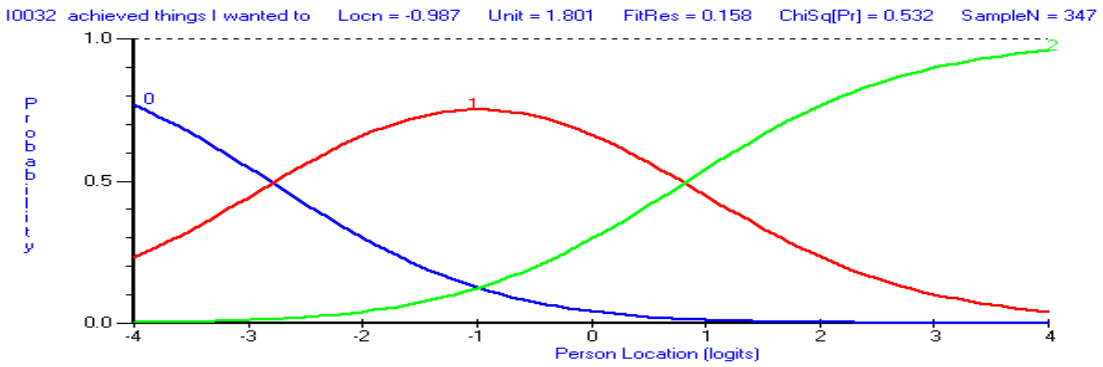


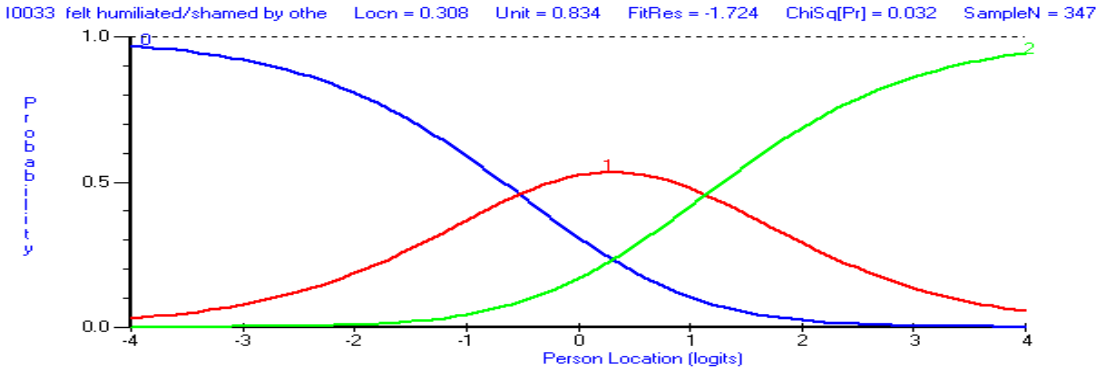
I0030 thought I am to blame for my p Locn = -0.636 Unit = 0.227 FitRes = 6.074 ChiSq[Pr] = 0.000 SampleN = 347

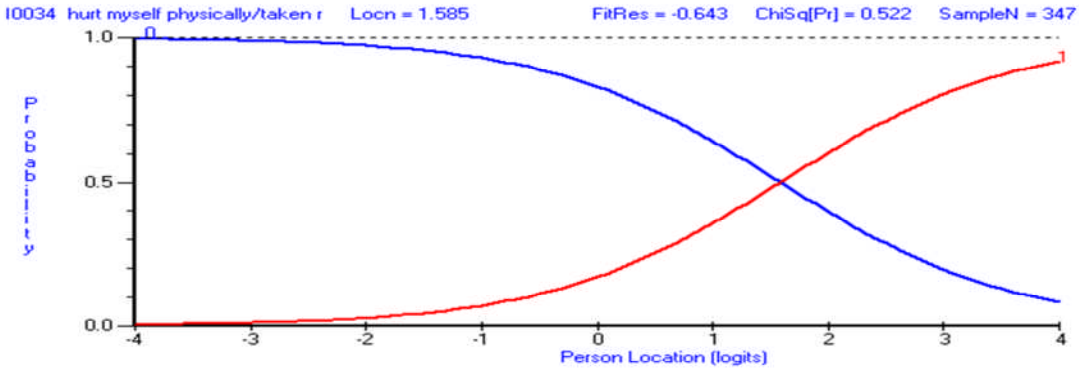


I0031 optimistic about future Locn = -0.640 Unit = 0.706 FitRes = 3.613 ChiSq[Pr] = 0.002 SampleN = 347




I0032 achieved things I wanted to Locn = -0.987 Unit = 1.801 FitRes = 0.158 ChiSq[Pr] = 0.532 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)


I0033 felt humiliated/shamed by othe Locn = 0.308 Unit = 0.834 FitRes = -1.724 ChiSq[Pr] = 0.032 SampleN = 347
Probability
1.0
0.5
0.0
0
1
2
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)


I0034 hurt myself physically/taken r Locn = 1.585 FitRes = -0.643 ChiSq[Pr] = 0.522 SampleN = 347
Probability
1.0
0.5
0.0
0
1
-4 -3 -2 -1 0 1 2 3 4
Person Location (logits)

Appendix 8. Validation of the emotional component of CORE-6D. Rasch analysis on random samples [N400b], [N1500] and [N1500v]

Validation on random sample [N400b]

Table A12. Item-person and item-trait interaction summary statistics – [N400b]

	Items		Persons	
	Location	**Fit residual**	**Location**	**Fit residual**
Mean	0.000	0.227	-0.678	-0.259
Standard deviation	1.312	0.914	1.422	0.775

Item-trait interaction		**Reliability indices**	
Total Item chi-square	22.970	**PSI**	0.65905
Total degrees of freedom	20	**Cronbach Alpha**	N/A (missing data)
Total chi-square probability	0.2903	**Power of test-of-fit**	Good

Table A13. Individual item fit statistics– [N400b]

Item	CD	Location	Residual	Chi-square	P-value
1. [terribly alone and isolated]	FC	-1.552	-0.363	3.799	0.434
15. [felt panic or terror]	SA	-0.727	-0.163	6.074	0.194
16. [made plans to end my life]	RS	1.932	-0.733	7.462	0.113
21. [able to do most things I needed to]	FG	0.454	1.328	4.861	0.302
33. [felt humiliated or shamed by other people]	FS	-0.106	1.067	0.774	0.942

Residuals ≥ $\left|2.5\right|$ are considered high; chi-square probabilities have been assessed using Bonferroni adjustment. CD = conceptual domain; FC = functioning-close relationships; FG = functioning-general; FS = functioning-social relationships; RS = risk/harm-to-self; SA = symptoms-anxiety

Table A14. Class interval distribution– [N400b]

ITEM	CI1	CI2	CI3	CI4	CI5	CI6	CI7
1	34	54	66	73	52	44	54
15	34	54	66	75	53	44	54
16	34	53	65	74	51	44	54
21	33	54	66	74	53	44	53
33	34	52	67	71	52	44	54

CI: Class interval

Figure A3. Rach item threshold map– [N400b]

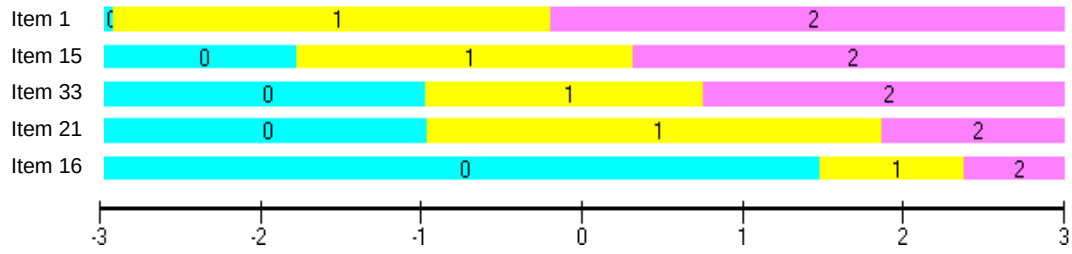


Figure A4. Item map – [N400b]

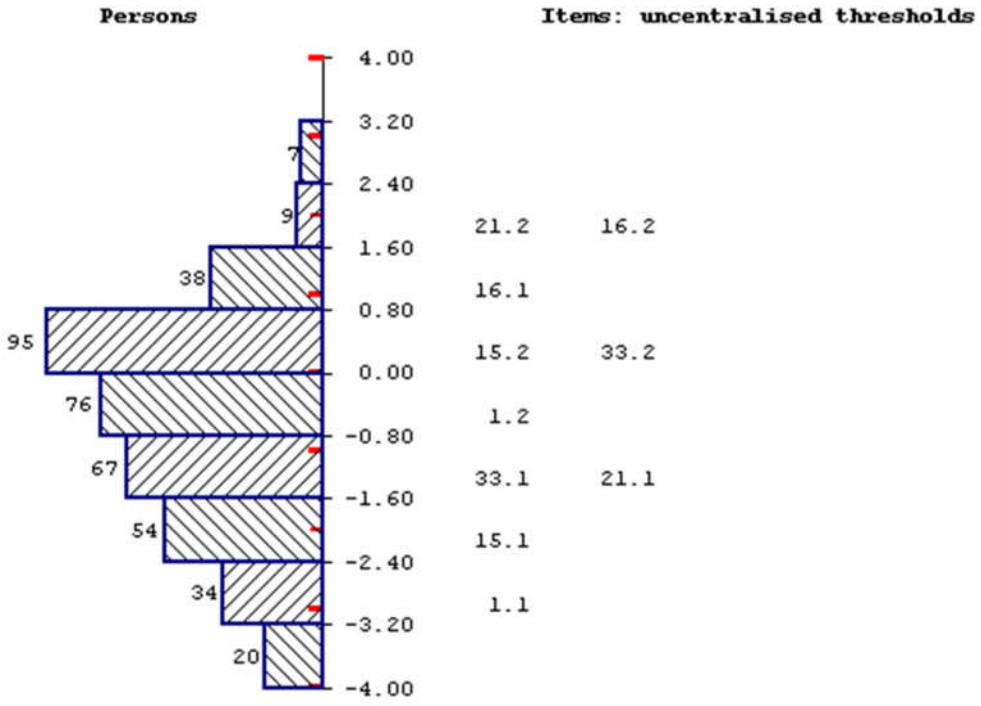


Table A15. Principal Component Analysis of item fit residuals – [N400b]

Item	Component 1	Component 2	Component 3	Component 4	Component 5
1	0.15	-0.19	**0.95**	-0.12	-0.18
15	0.18	**0.92**	-0.21	-0.14	-0.23
16	0.10	-0.11	-0.11	**0.98**	-0.09
21	**-0.93**	-0.18	-0.16	-0.13	-0.25
33	0.30	-0.26	-0.23	-0.13	**0.88**

Varimax rotation; loadings ≥ | 0.40 | are shown in bold

Table A16. Residual correlation matrix – [N400b]

Item	1	15	16	21	33
1	1.00				
15	-0.29	1.00			
16	-0.17	-0.17	1.00		
21	-0.19	-0.23	-0.16	1.00	
33	-0.27	-0.33	-0.12	-0.39	1.00

None of the correlation coefficients between different items is ≥ |0.40|

According to the post-hoc test of unidimensionality, the proportion of independent t-tests that were significant at the 0.05 level was 1.85%, which verified the unidimensionality of the emotional component of CORE-6D also in the random sample [N400b].

Validation on dataset [N1500]

Table A17. Item-person and item-trait interaction summary statistics after adjustment of sample size to n=500 – [N1500]

	Items		Persons	
	Location	**Fit residual**	**Location**	**Fit residual**
Mean	0.000	-0.035	-0.757	-0.246
Standard deviation	1.276	1.489	1.409	0.739

Item-trait interaction		**Reliability indices**	
Total Item chi-square	32.475	**PSI**	0.65990
Total degrees of freedom	35	**Cronbach Alpha**	N/A (missing data)
Total chi-square probability	0.5906	**Power of test-of-fit**	Good

Table A18. Individual item fit statistics after adjustment of sample size to n=500 – [N1500]

Item	CD	Location	Residual	Chi-square	P-value
1. [terribly alone and isolated]	FC	-1.514	-1.212	6.761	0.454
15. [felt panic or terror]	SA	-0.739	-0.245	6.078	0.531
16. [made plans to end my life]	RS	1.828	-1.676	8.059	0.327
21. [able to do most things I needed to]	FG	0.560	1.773	8.944	0.257
33. [felt humiliated or shamed by other people]	FS	-0.136	1.182	2.634	0.917

Residuals ≥ |2.5| are considered high; chi-square probabilities have been assessed using Bonferroni adjustment. CD = conceptual domain; FC = functioning-close relationships; FG = functioning-general; FS = functioning-social relationships; RS = risk/harm-to-self; SA = symptoms-anxiety

Table A19. Class interval distribution – [N1500]

ITEM	CI1	CI2	CI3	CI4	CI5	CI6	CI7
1	142	172	218	230	186	149	139
15	141	172	218	232	185	150	139
16	142	170	218	229	184	146	139
21	141	168	218	231	186	150	138
33	141	170	218	226	185	148	139

CI: Class interval

Figure A5. Rasch item threshold map – [N1500]

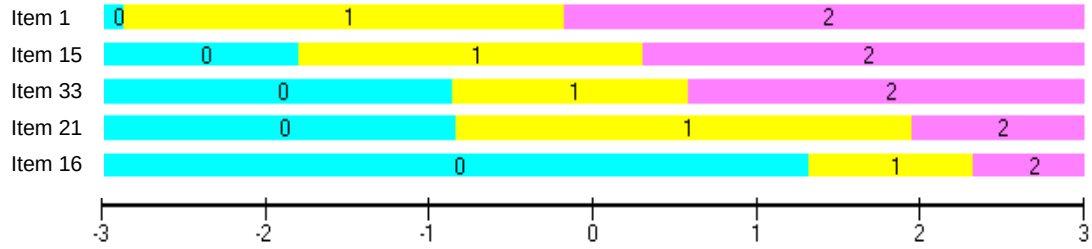


Figure A6. Item map – [N1500]

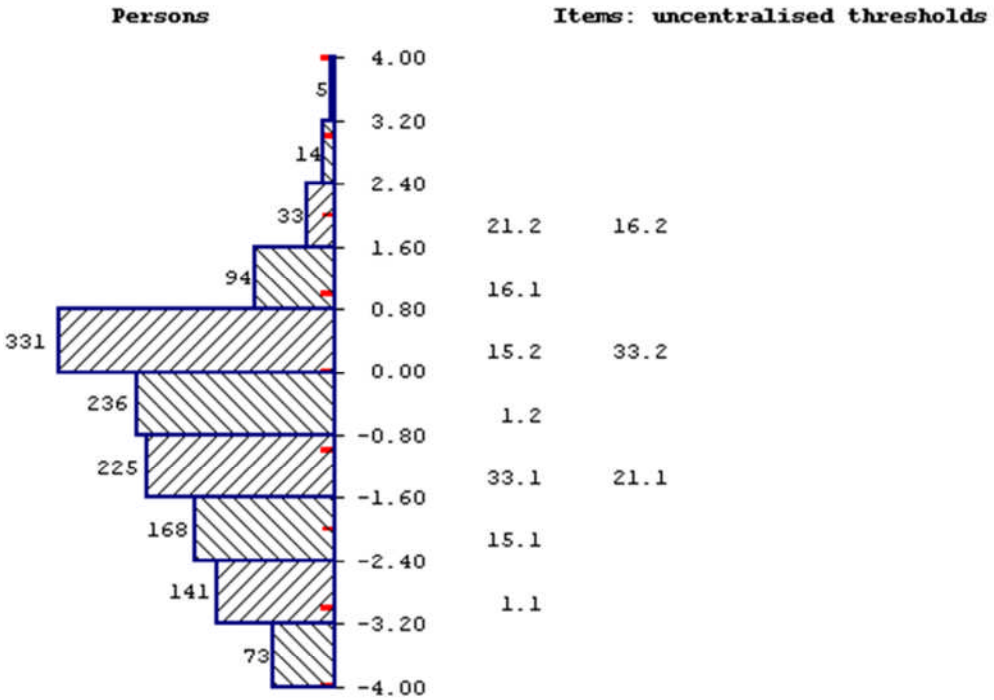


Table A20. Principal Component Analysis of the item fit residuals – [N1500]

Item	Component 1	Component 2	Component 3	Component 4	Component 5
1	-0.20	**0.94**	-0.18	-0.11	-0.19
15	-0.24	-0.20	-0.19	-0.11	**0.93**
16	-0.13	-0.10	-0.12	**0.98**	-0.09
21	-0.21	-0.18	**0.93**	-0.14	-0.18
33	**0.88**	-0.24	-0.25	-0.17	-0.28

Varimax rotation; loadings ≥ |0.40| are shown in bold

Table A21. Residual correlation matrix – [N1500]

Item	1	15	16	21	33
1	1.00				
15	-0.27	1.00			
16	-0.14	-0.12	1.00		
21	-0.24	-0.24	-0.19	1.00	
33	-0.28	-0.35	-0.20	-0.30	1.00

None of the correlation coefficients between different items is ≥ |0.40|

The post-hoc test of unidimensionality confirmed once again the unidimensionality of the measure, with the proportion of significant independent t-tests at the 0.05 level reaching 1.45%.

Validation on dataset [N1500v]

Table A22. Item-person and item-trait interaction summary statistics after adjustment of sample size to n=700 – [N1500v]

	Items		Persons	
	Location	**Fit residual**	**Location**	**Fit residual**
Mean	0.000	0.244	-0.816	-0.244
Standard deviation	1.173	1.298	1.407	0.722

Item-trait interaction		**Reliability indices**	
Total Item chi-square	34.745	**PSI**	0.6724
Total degrees of freedom	30	**Cronbach Alpha**	N/A (missing data)
Total chi-square probability	0.2521	**Power of test-of-fit**	Good

Table A23. Individual item statistics after adjustment of sample size to n=70 – [N1500v]

Item	CD	Location	Residual	Chi-square	P-value
1. [terribly alone and isolated]	FC	-1.422	-0.832	9.073	0.169
15. [felt panic or terror]	SA	-0.699	0.412	5.527	0.478
16. [made plans to end my life]	RS	1.594	-1.324	8.052	0.234
21. [able to do most things I needed to]	FG	0.661	1.615	6.984	0.322
33. [felt humiliated or shamed by other people]	FS	-0.134	1.349	5.108	0.530

Residuals ≥ |2.5| are considered high; chi-square probabilities have been assessed using Bonferroni adjustment. CD = conceptual domain; FC = functioning-close relationships; FG = functioning-general; FS = functioning-social relationships; RS = risk/harm-to-self; SA = symptoms-anxiety

Table A24. Class interval distribution – [N1500v]

ITEM	CI1	CI2	CI3	CI4	CI5	CI6	CI7
1	164	214	257	227	202	149	172
15	165	214	256	230	202	150	171
16	164	214	256	228	199	143	171
21	163	214	253	227	201	148	171
33	164	214	256	228	202	146	171

CI: Class interval

Figure A7. Rasch item threshold map – [N1500v]

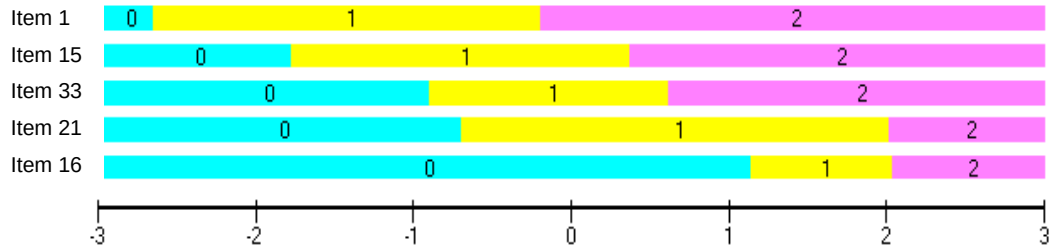


Figure A8. Item map - [N1500v]

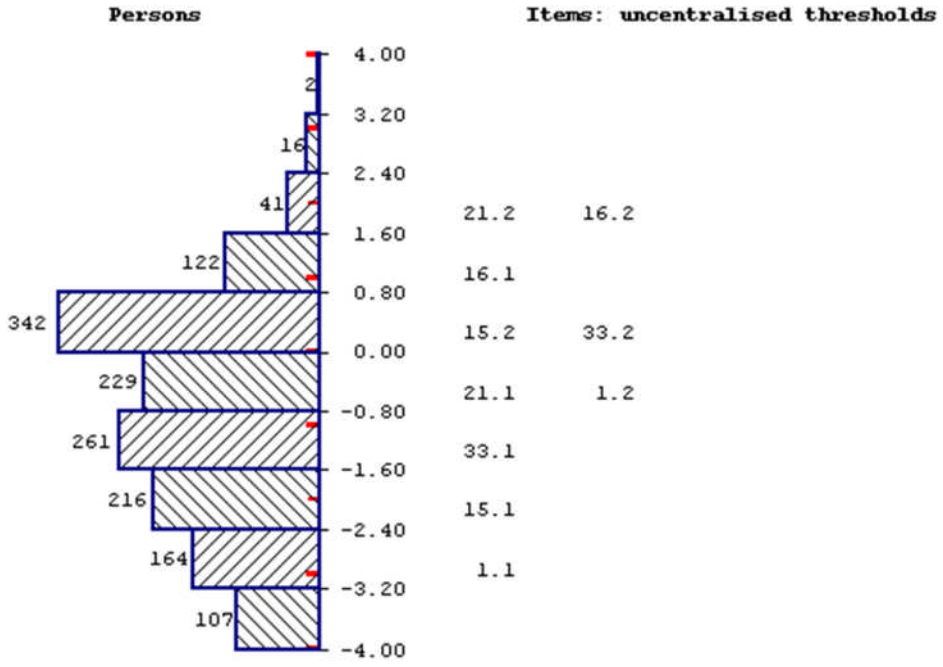


Table A25. Principal Component Analysis of item residuals – [N1500v]

Item	Component 1	Component 2	Component 3	Component 4	Component 5
1	-0.19	0.21	-0.16	-0.11	**0.94**
15	-0.24	**-0.91**	-0.18	-0.15	-0.23
16	-0.13	0.12	-0.13	**0.97**	-0.10
21	-0.21	0.17	**0.94**	-0.14	-0.15
33	**0.90**	0.25	-0.24	-0.16	-0.22

Varimax rotation; loadings ≥ |0.40| are shown in bold

Table A26. Residual correlation matrix – [N1500v]

Item	1	15	16	21	33
1	1.00				
15	-0.32	1.00			
16	-0.13	-0.18	1.00		
21	-0.20	-0.22	-0.20	1.00	
33	-0.27	-0.32	-0.19	-0.32	1.00

None of the correlation coefficients between different items is ≥ |0.40|

The post-hoc test of unidimensionality confirmed the unidimensionality of the new scale on [N1500v], with the proportion of significant independent t-tests at the 0.05 level reaching 0.79%.

Appendix 9. Interviewer booklet used in the valuation survey of CORE-6D

University of Sheffield - School of Health and Related Research

Survey of General Health Values
Interviewer booklet

Thank you very much for agreeing to take part in this survey. As we explained in the letter, this is a survey for the University of Sheffield about the way people value common mental and physical health problems.
You can stop the interview at any time if you do not want to continue. If there are any questions you do not want to answer tell the interviewer and they will move onto the next question.
All information you provide is confidential. The information you give will not be used in any way that could identify you.
We are interested in people's views, and there are no right or wrong answers. Please tell us what you think.

Respondent ID _____

Interviewer ID _____

Card bloc _____

Start time _____

End time _____

SECTION [A] SELF REPORTED HEALTH

BEFORE THE INTERVIEW PLEASE PREPARE THE BLOC OF CARDS THAT ARE BEING USED.

FILL IN RESPONDENT ID, INTERVIEWER ID, CARD BLOC AND START TIME ON FRONT PAGE OF THE INTERVIEWER BOOKLET AND SELF-COMPLETION BOOKLET.

READ ALOUD THE INTRODUCTION ON THE FRONT PAGE.

To start off, I would like you to answer a few questions about your own health and wellbeing.

There is one question about your life satisfaction. Then there are some statements about different aspects of your general health and quality of life. The statements are arranged in groups. For each group, please tick just one statement that best describes your own health state today.

HAND RESPONDENT SELF-COMPLETION BOOKLET

Please answer the questions in section A1 and A2 and then return the booklet to me.

AFTER SECTIONS A1 AND A2 HAVE BEEN COMPLETED TAKE THE SELF-COMPLETION BOOKLET FROM RESPONDENT.

MAKE SURE RESPONDENT HAS ONLY TICKED ONE BOX IN EACH GROUP.

HAND RESPONDENT SELF-COMPLETION BOOKLET OPEN AT SECTION A3, PAGE 3.

CARD BLOC 1 ONLY: Please answer the questions in section A3 and then return the booklet to me.

CARD BLOCS 2 AND 3 ONLY: Please answer the questions in section A3 and A4 and then return the booklet to me.

AFTER THE SECTION(S) HAS BEEN COMPLETED TAKE THE SELF-COMPLETION BOOKLET FROM RESPONDENT.

THE RESPONDENT CAN CHOOSE WHETHER TO ANSWER THESE QUESTIONS SO DO NOT CHECK THESE RESPONSES.

GO TO SECTION [B] RANKING EXERCISE

SECTION [B] RANKING EXERCISE

CARD BLOC 1: SHUFFLE 4 GREEN CARDS: **ST, ZB, MC, GQ**

CARD BLOCS 2 AND 3: SHUFFLE ALL 8 GREEN CARDS AND PICK 4 CARDS OUT AND PUT TO ONE SIDE FOR USE IN SECTIONS D AND E.

PLEASE SHUFFLE THE REMAINING 4 GREEN CARDS AND THE PINK CARD (FULL HEALTH) AND BLUE CARD (DEAD). DO NOT INCLUDE THE CREAM PRACTICE CARD.

THERE ARE 6 CARDS IN TOTAL.
SHOW PACK OF CARDS TO RESPONDENT

The booklet that you have just completed had questions made up from statements about health for you to choose from. I now have some cards which describe different states that you might find yourself in, and each of these is made up by combining the statements that you have just seen.

For example here is a card which has a description of a state written on it.

GIVE A GREEN CARD TO THE RESPONDENT TO LOOK AT.

If you were living in this state you would… READ CARD ALOUD

FOR EXAMPLE, for the card pictured below:

- You feel terribly alone and isolated often, most or all the time
- You feel panic or terror only occasionally or sometimes
- You feel humiliated or shamed by other people only occasionally or sometimes
- You are able to do most things you need to only occasionally or sometimes
- You never make plans to end your life

YOU WOULD SAY ALOUD: IF YOU WERE LIVING IN THIS STATE YOU WOULD FEEL TERRIBLY ALONE AND ISOLATED OFTEN, MOST OR ALL THE TIME, FEEL PANIC OR TERROR ONLY OCCASIONALLY OR SOMETIMES, FEEL HUMILIATED OR SHAMED BY OTHER PEOPLE ONLY OCCASIONALLY OR SOMETIMES, ARE ABLE TO DO MOST THINGS YOU NEED TO ONLY OCCASIONALLY OR SOMETIMES, NEVER MAKE PLANS TO END YOUR LIFE.

We are now going to use a technique called ranking to find out how good or bad you think living in some of the states would be. The states that we will show you have nothing to do with the answers you have just provided about your own health and well-being.

Now, here is a set of 6 cards. Each of them has a description of a state written on it. Each card has a different state description on it.

I would like you to place the cards in order of how good or bad you think they are. I would like you to imagine that you yourself are actually in each state and that you would live in that state for 10 years. After this time period you must assume that you would die. Please read each card carefully to see exactly what the state is and how it differs from the others. When you have finished reading through, please place the cards in order of how good or bad you think they are. Put the one you think is best at

the top (POINT TO TOP END), and the one that you think is worst at the bottom (POINT TO THE BOTTOM END).

PASS CARDS TO RESPONDENT

If you think two states are equal, put them side by side. You will notice that there is a card which says “dead”. Please also put this with the other cards in order where you think it belongs. You can change your ordering at any time.

RECORD THE RESULTS OF THE RANKING EXERCISE IN THE TABLE BELOW. IF MORE THAN TWO CARDS ARE RANKED EQUALLY, CROSS OUT THE NUMBER IN THE RANK COLUMN AND WRITE THE CORRECT RANK.

RANK	CARD CODE e.g. WG
1	
2	
3	
4	
5	
6	

SECTION [C] VALUATION EXERCISE

C. INTERVIEWER SCRIPT FOR TTO

REMOVE PINK CARD AND BLUE CARD FROM THE PACK OF RANKED CARDS.

PLEASE SHUFFLE ALL 4 GREEN CARDS (THE 4 CARDS THAT HAVE JUST BEEN RANKED).
PICK UP PINK AND BLUE CARDS, AND CREAM PRACTICE CARD

HAVE TTO BOARD SIDE '1' FACING UPWARDS.
SET BOARD MARKER FOR LIFE A TO 10 YEARS.

Now we are going to use a technique called the time trade off to find out how good or bad you think living in some of the health states would be. The time trade off asks you to compare living in two health states for a maximum period of 10 years. After this time period you must assume that you would die.

I'm going to start with a practice using a health state which is similar to those which you have just ranked.
I am going to ask you to make a choice between living in this health state (Life B – cream card) and living in another health state (Life A – pink card). The pink scale and the green scale show the number of years you would be in each state for. Remember, I want you to imagine that <u>you</u> are in these states.

C2a. **INTERVIEWER CHECK:**

PICK OUT CREAM PRACTICE TTO CARD.
TICK TO CONFIRM CORRECT CARD: _____

PASS CARD TO THE RESPONDENT.

Please read this card carefully.

b. PLACE CREAM PRACTICE TTO CARD IN POCKET FOR LIFE B.
PLACE PINK CARD IN POCKET FOR LIFE A.
MAKE SURE THAT BOARD MARKER FOR LIFE A IS AT 10 YEARS.

At the moment, each scale says 10 years. This means that you would either live in Life A for 10 years and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

Life A	1. GO TO C3
Life B	2. ASK c.
The same	3.

c. IF 'LIFE B' AT b.: Does this mean that you would rather live in Life B for 10 years than in Life A for 10 years?

IF 'THE SAME' AT b.: Does this mean that living in Life B for 10 years would be the same as living in Life A for 10 years?

Yes	1 GO TO C3
No (first time)	2 Repeat b
No (second time)	3 GO TO C3

C3a. CONTINUE WORKING WITH CREAM PRACTICE CARD.
TICK TO CONFIRM CORRECT CARD: _____

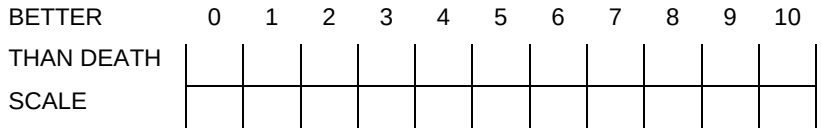
b. MOVE BOARD MARKER FOR LIFE A TO 0 YEARS.
Now you would either die immediately, or you would live in Life B for 10 years and then die. Would you prefer to die immediately or to have Life B, or are they the same?

Life A	1. GO TO h. (STATE WORSE THAN DEATH)
Life B	2. GO TO c. (STATE BETTER THAN DEATH)
The same	3. GO TO C4

ASK IF 'LIFE B' (code 2) AT b.

c. <u>STATE BETTER THAN DEATH</u>

MARK 'X' UNDER 0 ON THE SCALE BELOW.



CONTINUE TO USE TIME BOARD WITH SIDE '1' UPWARDS

SET BOARD MARKER FOR LIFE A TO 5 YEARS (t=5).

d. Now you would either live in Life A for 't' years and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

CONTINUE TO WRITE ON SCALE <u>ABOVE ON THIS PAGE</u>.

IF A: ✓ UNDER 't' MOVE MARKER 1 YEAR TO THE LEFT.
REPEAT d. WITH 't' 1 LESS THAN LAST TIME.

IF B: × UNDER 't' MOVE MARKER 1 YEAR TO THE RIGHT.
REPEAT d. WITH 't' 1 MORE THAN LAST TIME

IF SAME: = UNDER 't' GO TO C4

REPEAT d. UNTIL:

A) YOU ENTER '=' GO TO C4 <u>OR</u>

B) '×' AND '✓' APPEAR NEXT TO EACH OTHER GO TO e.

ASK IF d. ENDED WITH '×' AND '✓' NEXT TO EACH OTHER

e. LET 't' NOW BE HALFWAY BETWEEN THE ADJACENT CROSS AND TICK, I.E. 'SOMETHING AND 6 MONTHS'

What if you would either live in Life A for 't' and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

Life A	1. GO TO C4
Life B	2. GO TO f.
The same	3. GO TO C4

f. IF 'LIFE B' (code 2) AT e.

INTERVIEWER CHECK:

IF THERE IS A × UNDER 9	1. GO TO g.
IF THERE IS <u>NOT</u> A × UNDER 9	2. GO TO C4

ASK IF THERE IS '×' UNDER 9 AND '✓' UNDER 10

g. Would you be prepared to sacrifice <u>any</u> time in order to avoid Life B?

IF YES: How many weeks?

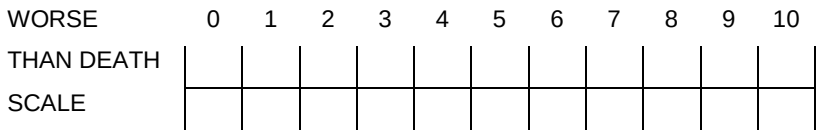
ENTER WEEKS: _____

Yes	1. ___	GO TO C4
No	2.	

ASK IF 'LIFE A' (code 1) AT b.

h. <u>STATE WORSE THAN DEATH</u>

MARK '✓' UNDER 0 ON SCALE BELOW.



TURN TTO BOARD SIDE '2' UPWARDS.

MOVE CREAM PRACTICE CARD TO TOP LEFT POCKET ON SIDE '2'.

PLACE PINK CARD IN TOP RIGHT POCKET ON SIDE '2'.

PLACE BLUE CARD IN BOTTOM POCKET ON SIDE '2'.

SET BOARD MARKER FOR LIFE A TO 5 YEARS (t = 5).

Now here is a different choice.

i. Life A is now 't' years of this state (POINT TO THE CREAM PRACTICE CARD) followed by '10-t' years in this other state (POINT TO THE PINK CARD). Or instead of that you could choose to die immediately (POINT TO LIFE B). Would you prefer Life A, or to die immediately, or are they the same?
WRITE ON SCALE ABOVE ON THIS PAGE.

IF A:	✓ UNDER 't'	MOVE MARKER 1 YEAR TO THE RIGHT. REPEAT i. WITH 't' 1 MORE THAN LAST TIME.
IF B:	× UNDER 't'	MOVE MARKER 1 YEAR TO THE LEFT. REPEAT i. WITH 't' 1 LESS THAN LAST TIME.
IF SAME:	= UNDER 't'	GO TO C4

REPEAT i. UNTIL:

A) YOU ENTER '='	GO TO C4 OR
B) '✓' AND '×' APPEAR NEXT TO EACH OTHER	GO TO j.

ASK IF i. ENDED WITH '✓' AND '×' NEXT TO EACH OTHER

j. LET 't' NOW BE HALFWAY BETWEEN THE ADJACENT TICK AND CROSS, I.E. 'SOMETHING AND 6 MONTHS'.
What if Life A was 't' of this state (POINT TO THE CREAM PRACTICE CARD) followed by '10-t' in this other state (POINT TO THE PINK CARD). Or instead of that you could choose to die immediately (POINT TO LIFE B). Would you prefer Life A, or to die immediately, or are they the same?

Life A	1.
Life B	2. GO TO C4
The same	3.

C4a. **INTERVIEWER CHECK:**

PICK UP PACK OF 4 GREEN HEALTH STATE CARDS (SHUFFLED) THAT WERE JUST RANKED.

TAKE OUT FIRST CARD TO BE VALUED. ENTER LETTERS OF THE CARD: _____

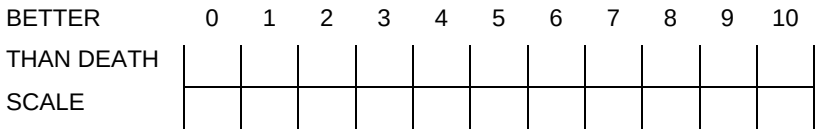
PASS CARD TO THE RESPONDENT.
Please read this card through carefully.

b. HAVE TTO BOARD WITH SIDE '1' FACING UPWARDS.
PLACE GREEN CARD IN POCKET FOR LIFE B.
MOVE BOARD MARKER FOR LIFE A TO 0 YEARS.
Now you would either die immediately, or you would live in Life B for 10 years and then die. Would you prefer to die immediately or to have Life B, or are they the same?

Life A	1. GO TO h. (STATE WORSE THAN DEATH)
Life B	2. GO TO c. (STATE BETTER THAN DEATH)
The same	3. GO TO C5

ASK IF 'LIFE B' (code 2) AT b.

c. <u>STATE BETTER THAN DEATH</u>
MARK 'X' UNDER 0 ON THE SCALE BELOW.



CONTINUE TO USE TIME BOARD WITH SIDE '1' UPWARDS
SET BOARD MARKER FOR LIFE A TO 5 YEARS (t=5).

d. Now you would either live in Life A for 't' years and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

CONTINUE TO WRITE ON SCALE ABOVE ON THIS PAGE.

IF A:	✓ UNDER 't'	MOVE MARKER 1 YEAR TO THE LEFT. REPEAT d. WITH 't' 1 LESS THAN LAST TIME.
IF B:	× UNDER 't'	MOVE MARKER 1 YEAR TO THE RIGHT. REPEAT d. WITH 't' 1 MORE THAN LAST TIME.
IF SAME:	= UNDER 't'	GO TO C5

REPEAT d. UNTIL:

A) YOU ENTER '='	GO TO C5 OR
B) '×' AND '✓' APPEAR NEXT TO EACH OTHER	GO TO e.

ASK IF d. ENDED WITH '×' AND '✓' NEXT TO EACH OTHER

e. LET 't' NOW BE HALFWAY BETWEEN THE ADJACENT CROSS AND TICK, I.E. 'SOMETHING AND 6 MONTHS'

What if you would either live in Life A for 't' and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

Life A	1. GO TO C5
Life B	2. GO TO f.
The same	3. GO TO C5

IF 'LIFE B' (code 2) AT e.

f. INTERVIEWER CHECK:

IF THERE IS A × UNDER 9	1. GO TO g.
IF THERE IS NOT A × UNDER 9	2. GO TO C5

ASK IF THERE IS '×' UNDER 9 AND '✓' UNDER 10

g. Would you be prepared to sacrifice <u>any</u> time in order to avoid Life B?

IF YES: How many weeks?

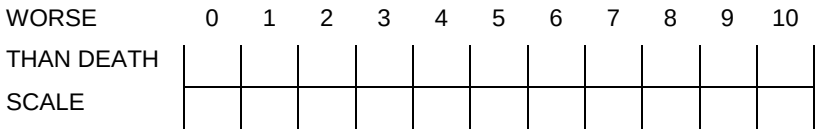
ENTER WEEKS: _____

Yes	1. ___ GO TO C5
No	2.

ASK IF 'LIFE A' (code 1) AT b.

h. <u>STATE WORSE THAN DEATH</u>

MARK '✓' UNDER 0 ON SCALE BELOW.



TURN TTO BOARD SIDE '2' UPWARDS.

MOVE GREEN CARD TO TOP LEFT POCKET ON SIDE '2'.

SET BOARD MARKER FOR LIFE A TO 5 YEARS (t = 5).

Now here is a different choice.

i. Life A is now 't' years of this state (POINT TO THE GREEN CARD) followed by '10-t' years in this other state (POINT TO THE PINK CARD). Or instead of that you could choose to die immediately (POINT TO LIFE B). Would you prefer Life A, or to die immediately, or are they the same?

WRITE ON SCALE ABOVE <u>ON THIS PAGE</u>.

IF A:	✓ UNDER 't'	MOVE MARKER 1 YEAR TO THE RIGHT. REPEAT i. WITH 't' 1 MORE THAN LAST TIME.
IF B:	× UNDER 't'	MOVE MARKER 1 YEAR TO THE LEFT. REPEAT i. WITH 't' 1 LESS THAN LAST TIME.
IF SAME:	= UNDER 't'	GO TO C5

REPEAT i UNTIL:

A) YOU ENTER '=' GO TO C5 <u>OR</u>

B) '✓' AND '×' APPEAR NEXT TO EACH OTHER GO TO j.

ASK IF i. ENDED WITH '✓' AND '×' NEXT TO EACH OTHER

j. LET 't' NOW BE HALFWAY BETWEEN THE ADJACENT TICK AND CROSS, I.E. 'SOMETHING AND 6 MONTHS'.

What if Life A was 't' of this state (POINT TO THE GREEN CARD) followed by '10-t' in this other state (POINT TO THE PINK CARD). Or instead of that you could choose to die immediately (POINT TO LIFE B). Would you prefer Life A, or to die immediately, or are they the same?

Life A	1.
Life B	2. GO TO C5
The same	3.

C5a. **TAKE OUT SECOND CARD TO BE VALUED. ENTER LETTERS OF THE CARD: _____**

[process repeated as above – where 'GO TO C5' replace by 'GO TO C6']

C6a. **TAKE OUT THIRD CARD TO BE VALUED. ENTER LETTERS OF THE CARD: _____**

[process repeated as above – where 'GO TO C5' replace by 'GO TO C7']

C7a. **TAKE OUT FOURTH CARD TO BE VALUED. ENTER LETTERS OF THE CARD: _____**

[process repeated as above – where 'GO TO C5' replace by 'GO TO SECTION [D]']

SECTION [D] SECOND RANKING EXERCISE

CARD BLOC 1 ONLY: SHUFFLE REMAINING 4 GREEN CARDS **KX, RL, WA, NV**

HAND RESPONDENT SELF-COMPLETION BOOKLET OPEN AT SECTION A4, PAGE 4.

Please answer the question in section A4 and then return the booklet to me.

AFTER SECTION A4 HAS BEEN COMPLETED TAKE THE SELF-COMPLETION BOOKLET FROM RESPONDENT.

PLEASE SHUFFLE THE 4 GREEN CARDS AND THE PINK CARD (FULL HEALTH) AND BLUE CARD (DEAD). DO NOT INCLUDE THE CREAM PRACTICE CARD.

THERE ARE 6 CARDS IN TOTAL.
SHOW PACK OF CARDS TO RESPONDENT

We are now going to repeat the exercises you have just done for a different set of cards. First we are going to use a technique called ranking to find out how good or bad you think living in some of the states would be. Again, the states that we will show you have nothing to do with the answers you have just provided about your own health and well-being.

Now, here is a set of 6 cards. Each of them has a description of a state written on it. Each card has a <u>different</u> state description on it.

I would like you to place the cards in order of how good or bad you think they are. I would like you to imagine that <u>you yourself</u> are actually in each state and that you would live in that state <u>for 10 years</u>. After this time period you must assume that you would die. Please read each card carefully to see exactly what the state is and how it differs from the others. When you have finished reading through, please place the cards in order of how good or bad you think they are. Put the one you think is best at the top (POINT TO TOP END), and the one that you think is worst at the bottom (POINT TO THE BOTTOM END).

PASS CARDS TO RESPONDENT

If you think two states are equal, put them side by side. You will notice that there is a card which says “dead”. Please also put this with the other cards in order where you think it belongs. You can change your ordering at any time.

RECORD THE RESULTS OF THE RANKING EXERCISE IN THE TABLE BELOW. IF MORE THAN TWO CARDS ARE RANKED EQUALLY, CROSS OUT THE NUMBER IN THE RANK COLUMN AND WRITE THE CORRECT RANK.

RANK	CARD CODE e.g. WG
1	
2	
3	
4	
5	
6	

SECTION [E] SECOND VALUATION EXERCISE

E. INTERVIEWER SCRIPT FOR TTO

REMOVE PINK CARD AND BLUE CARD FROM THE PACK OF RANKED CARDS.

PLEASE SHUFFLE ALL 4 GREEN CARDS (THAT HAVE JUST BEEN RANKED) AND PLACE ON ONE SIDE.
PICK UP PINK AND BLUE CARDS.

HAVE TTO BOARD SIDE '1' FACING UPWARDS.
SET BOARD MARKER FOR LIFE A TO 10 YEARS.

Now we are going to use a technique called the time trade off to find out how good or bad you think living in some of the health states would be. The time trade off asks you to compare living in two health states for a maximum period of 10 years. After this time period you must assume that you would die.

I am going to ask you to make a choice between living in this health state (Life B – green card) and living in another health state (Life A – pink card). The pink scale and the green scale show the number of years you would be in each state for. Remember, I want you to imagine that <u>you</u> are in these states.

E2a. **INTERVIEWER CHECK:**

PICK OUT FIFTH CARD TO BE VALUED. ENTER LETTERS OF THE CARD:

_____

PASS CARD TO THE RESPONDENT.
Please read this card carefully.

b. PLACE GREEN TTO CARD IN POCKET FOR LIFE B.
PLACE PINK CARD IN POCKET FOR LIFE A.
MAKE SURE THAT BOARD MARKER FOR LIFE A IS AT 10 YEARS.

At the moment, each scale says 10 years. This means that you would either live in Life A for 10 years and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

Life A	1. GO TO E3
Life B	2. ASK c.
The same	3.

c. IF 'LIFE B' AT b.: Does this mean that you would rather live in Life B for 10 years than in Life A for 10 years?

IF 'THE SAME' AT b.: Does this mean that living in Life B for 10 years would be the same as living in Life A for 10 years?

Yes	1. GO TO E3
No (first time)	2. Repeat b
No (second time)	3. GO TO E3

E3a. **CONTINUE WORKING WITH GREEN CARD**
[process repeated as in section C4 – where 'GO TO C5' replace by 'GO TO E4']

E4a. **TAKE OUT SIXTH CARD TO BE VALUED. ENTER LETTERS OF THE CARD:** _____
[process repeated as in section C4 – where 'GO TO C5' replace by 'GO TO E5']

E5a. **TAKE OUT SEVENTH CARD TO BE VALUED. ENTER LETTERS OF THE CARD:** _____
[process repeated as in section C4 – where 'GO TO C5' replace by 'GO TO E6']

E6a. **TAKE OUT EIGHTH CARD TO BE VALUED. ENTER LETTERS OF THE CARD:** _____
[process repeated as in section C4 – where 'GO TO C5' replace by 'GO TO SECTION [F]']

SECTION [F] BACKGROUND CHARACTERISTICS

1. Some people have said that they found it quite difficult to answer the questions I have asked you. Others have said that they found it quite easy. How about you – did you find the ranking question – (READ ALOUD)

Very difficult……☐
Quite difficult……..☐
Neither difficult nor easy…....☐
Fairly easy……...☐
Or very easy ……...☐

2. How about you – did you find the time trade off questions – (READ ALOUD)

Very difficult……..☐
Quite difficult……...☐
Neither difficult nor easy…..….☐
Fairly easy……....☐
Or very easy ……...☐

3. Did you find the second ranking question –

Easier than the first ranking question……☐
More difficult than the first ranking question……..…☐
About the same……..☐

4. Did you find the second set of time trade off questions –

Easier than the first set of time trade off questions……..…☐
More difficult than the first set of time trade off questions……..…☐
About the same……..……..☐

I would now like you to answer some background questions about yourself. This information will help us to understand your answers better. If you have any comments to make about this interview, please feel free to use the last page of the booklet.

HAND RESPONDENT SELF-COMPLETION BOOKLET OPEN AT SECTION [F] BACKGROUND CHARACTERISTICS AND HEALTH SERVICE USE

AFTER SECTION [F] HAS BEEN COMPLETED TAKE THE SELF-COMPLETION BOOKLET FROM THE RESPONDENT

FILL IN “END TIME” ON FRONT PAGE OF INTERVIEWER SCRIPT
Thank you very much for your cooperation and your time.

SECTION [G] INTERVIEWER FEEDBACK TO BE COMPLETED AFTER THE INTERVIEW

1. How well do you think the respondent understood and carried out the first ranking exercise during the interview?

Understood and performed exercises easily……..... ☐
Some problems but seemed to understand the exercises in the end..... ☐
Doubtful whether the respondent understood the exercises…….. ☐

2. In terms of effort and concentration, which one of the following statements best describes the way the respondent undertook the first ranking exercise?

Concentrated very hard and put a great deal of effort into it……. ☐
Concentrated fairly hard and put some effort into it……... ☐
Didn't concentrate very hard and put little effort into it…… ☐
Concentrated at the beginning but lost interest/concentration before reaching the end…….. ☐

3. How well do you think the respondent understood and carried out the second ranking exercise during the interview?

Understood and performed exercises easily……..... ☐
Some problems but seemed to understand the exercises in the end..... ☐
Doubtful whether the respondent understood the exercises…….…….. ☐

4. In terms of effort and concentration, which one of the following statements best describes the way the respondent undertook the second ranking exercise?

Concentrated very hard and put a great deal of effort into it……. ☐
Concentrated fairly hard and put some effort into it……... ☐
Didn't concentrate very hard and put little effort into it…….. ☐
Concentrated at the beginning but lost interest/concentration before reaching the end…….. ☐

5. How well do you think the respondent understood and carried out the first set of time trade off exercises during the interview?

Understood and performed exercises easily……..... ☐
Some problems but seemed to understand the exercises in the end..... ☐
Doubtful whether the respondent understood the exercises…….. ☐

6. In terms of effort and concentration, which one of the following statements best describes the way the respondent undertook the first set of time trade off exercises?

Concentrated very hard and put a great deal of effort into it……☐
Concentrated fairly hard and put some effort into it……..☐
Didn't concentrate very hard and put little effort into it……☐
Concentrated at the beginning but lost interest/concentration before reaching the end……..…☐

7. How well do you think the respondent understood and carried out the second set of time trade off exercises during the interview?

Understood and performed exercises easily……...☐
Some problems but seemed to understand the exercises in the end…..☐
Doubtful whether the respondent understood the exercises……..☐

8. In terms of effort and concentration, which one of the following statements best describes the way the respondent undertook the second set of time trade off exercises?

Concentrated very hard and put a great deal of effort into it……☐
Concentrated fairly hard and put some effort into it……..☐
Didn't concentrate very hard and put little effort into it……☐
Concentrated at the beginning but lost interest/concentration before reaching the end……..…☐

Appendix 10. Ethical approval for the valuation survey of CORE-6D

The University Of Sheffield.

Cheryl Oliver
Ethics Committee Administrator

Regent Court
30 Regent Street

Telephone: +44 (0) 114 2220871
Fax: +44 (0) 114 272 4095 (non confidential)
Email: c.a.oliver@sheffield.ac.uk

Our ref: /CAO

30 March 09

Donna Rowen
ScHARR

Dear Donna

HTA Methodology project CoSMeQ (Condition-Specific Methodology for estimating QALYs): Developing and testing methods for deriving preference-based measures of health from condition specific measures (CoSMeQ)

Thank you for submitting the above research project for approval by the ScHARR Research Ethics Committee. On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that the project was approved.

If during the course of the project you need to deviate significantly from the documents you submitted for review, please inform me since written approval will be required.

Yours sincerely

Cheryl Oliver
Ethics Committee Administrator

Appendix 11. Self-completion booklet provided to participants in the valuation survey of CORE-6D

University of Sheffield - School of Health and Related Research

Survey of General Health Values
Self-completion booklet

This questionnaire contains questions which ask about various aspects of your health and about you. You may feel that some of these questions do not apply to you, but it is important that we ask everyone the same questions. Also a few questions are similar; please excuse the apparent overlap and try to answer each question independently.

Please read each question and consider your answers carefully. For each question, please read all answers and select one answer that best describes you. There are no right or wrong answers; what we want is your opinion.

You can stop the interview at any time if you do not want to continue. If there are any questions you do not want to answer leave them blank and move onto the next question.

Respondent ID _____

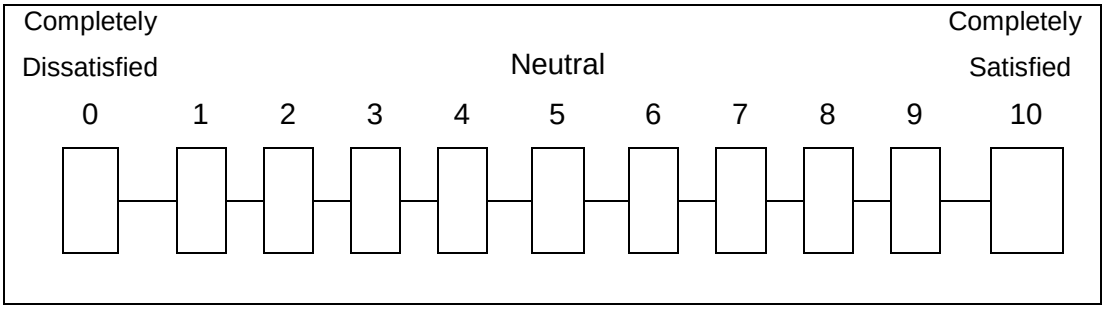
Interviewer ID _____

Card bloc _____

SECTION [A] YOUR HEALTH

[A1] Life satisfaction

1. Thinking about your own life and personal circumstances, how satisfied are you **with your life as a whole?**



[A2] Your own health in general

The following questions ask about your health in general. There are five groups of statements, each covering a different aspect of health. Please tick one statement in each group to show the statement which best describes your own health state TODAY.

Please tick one in each group

1. **Mobility**
   I have no problems in walking about……..….☐
   I have some problems in walking about……..….☐
   I am confined to bed……...☐

2. **Self-care**
   I have no problems with self-care..…….……☐
   I have some problems washing or dressing myself…....☐
   I am unable to wash or dress myself…….……...……...☐

3. **Usual activities**
   (e.g. work, study, housework, family or leisure activities)
   I have no problems with performing my usual activities…….…☐
   I have some problems with performing my usual activities……☐
   I am unable to perform my usual activities..……..…….....☐

4. **Pain and discomfort**
   I have no pain or discomfort..……..….....☐
   I have moderate pain or discomfort…....☐
   I have extreme pain or discomfort…….....☐

5. **Anxiety/Depression**
   I am not anxious or depressed..……...☐
   I am moderately anxious or depressed.…….....☐
   I am extremely anxious or depressed…….....☐

[A3] Your own health

The following questions ask about your health. There are five groups of statements. Please read each statement and think how often you felt that way last week. Please tick one statement in each group to show the statement which best describes your own health state DURING THE PAST WEEK.

1. Close relationships

I have never felt terribly alone and isolated.....☐

I have felt terribly alone and isolated only occasionally.....☐

I have felt terribly alone and isolated sometimes.....☐

I have felt terribly alone and isolated often.....☐

I have felt terribly alone and isolated most or all the time.....☐

2. Anxiety

I have never felt panic or terror…….....☐

I have felt panic or terror only occasionally…….....☐

I have felt panic or terror sometimes…….....☐

I have felt panic or terror often…….....☐

I have felt panic or terror most or all the time…….....☐

3. Social relationships

I have never felt humiliated or shamed by other people……..……☐

I have felt humiliated or shamed by other people only occasionally……☐

I have felt humiliated or shamed by other people sometimes…….....☐

I have felt humiliated or shamed by other people often…….....☐

I have felt humiliated or shamed by other people most or all the time…….....☐

4. Functioning

I have been able to do most things I need to most or all the time.....☐

I have been able to do most things I need to often…….....☐

I have been able to do most things I need to sometimes…….....☐

I have been able to do most things I need to only occasionally…….....☐

I have not been able to do most things I need to..……..…….....☐

5. Risk/Harm to self

I have never made plans to end my life.…….....☐

I have made plans to end my life only occasionally….....☐

I have made plans to end my life sometimes.…….....☐

I have made plans to end my life often..…….....☐

I have made plans to end my life most or all the time.…....☐

[A4] Your own health

The following question asks about your health. Please read each statement and think how often you felt that way last week. Please tick one statement to show the statement which best describes your own health state DURING THE PAST WEEK.

1. Physical health

I have never been troubled by aches, pains or other physical problems…..…….....☐

I have been troubled by aches, pains or other physical problems only occasionally……..…...☐

I have been troubled by aches, pains or other physical problems sometimes……☐

I have been troubled by aches, pains or other physical problems often.…….....☐

I have been troubled by aches, pains or other physical problems most or all the time.....☐

SECTION [F] Background characteristics and health service use

1. Are you:
   Male……☐
   Female…….☐

2. What is your age? _____

3. Are you:
   Single ……☐
   Married/Partner…….....☐
   Divorced/Separated…..☐
   Widowed…….....☐
   Not known…….....☐

4. Which of the following best describes your main activity?
   In employment or self-employment.....☐
   Retired……..…….…☐
   Housework……..……☐
   Student……...☐
   Seeking work/ Unemployed...…….……..☐
   Long-term sick……...☐
   Other (please specify)……☐ _____

5. Did your education continue after the minimum school leaving age?

   Yes…….……..☐
   No……...☐

6. Do you have a Degree or equivalent professional qualification?

   Yes…….……..☐
   No……...☐

7. How would you define your ethnic origin?
   White (British)……☐
   White (Irish)……..☐
   White (Other)……..…☐
   Mixed (White and Black Caribbean)..……☐
   Mixed (White and Black African)……..☐
   Mixed (White and Asian)…….....☐
   Mixed (other)……....☐
   Asian (Indian).....☐
   Asian (Pakistani)…….....☐
   Asian (Bangladeshi)…….....☐
   Asian (Chinese)..……..☐
   Asian (other)……..…….....☐
   Black (Caribbean)…….....☐
   Black (African) …….....☐
   Black (other) ……....☐
   Other……....☐

8. How often do you talk to any of your neighbours?
   On most days……...☐
   Once or twice a week….....☐
   Once or twice a month…....☐
   Less than once a month.....☐
   Never…….....☐
   Don't know…….....☐

9. How often do you meet friends or relatives who are not living with you?
   On most days……...☐
   Once or twice a week….....☐
   Once or twice a month……..☐
   Less than once a month.....☐
   Never…….....☐
   Don't know…….....☐

10. Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?
    Most people can be trusted…….....☐
    Can't be too careful in dealing with people……☐
    It depends on people/circumstances….....☐
    Don't know…….....☐

11. Have you experienced serious illness due to physical health?
    in you yourself……Yes ☐……No ☐
    in your family……...Yes ☐……No ☐
    in caring for others……..Yes ☐……No ☐

12. Have you experienced serious illness due to mental health?
    in you yourself……Yes ☐……No ☐
    in your family……...Yes ☐……No ☐
    in caring for others……..Yes ☐……No ☐

13. During the past 2 weeks, did you talk to a GP or family doctor on your own behalf for any reason at all?

    Yes……..☐
    No……....☐
    Don't know…..☐

14. Have you had any time off work due to your health in the last 4 weeks?

    Yes……..☐ → How many days? _____
    No……....☐
    Don't know…..☐

15. Does your household:
    own your home outright, or with a mortgage……☐
    rent from a local authority……..….☐
    rent from the private sector……....☐

If you have any general comments you would like to make about this interview please write them in the space provided below.

……
……
……
……
……
……

Thank you very much for your participation.

Please give the booklet to the interviewer.

Appendix 12. Plausible health states of the emotional component of CORE-6D, as identified by inspection of the Rasch item threshold map

Health state 1
I never feel terribly alone and isolated
I never feel panic or terror
I never feel humiliated or shamed by other people
I am able to do most things I need to often, most or all the time
I never make plans to end my life

Health state 2
I feel terribly alone and isolated only occasionally or sometimes
I never feel panic or terror
I never feel humiliated or shamed by other people
I am able to do most things I need to often, most or all the time
I never make plans to end my life

Health state 3
I feel terribly alone and isolated only occasionally or sometimes
I feel panic or terror only occasionally or sometimes
I never feel humiliated or shamed by other people
I am able to do most things I need to often, most or all the time
I never make plans to end my life

Health state 4
I feel terribly alone and isolated only occasionally or sometimes
I feel panic or terror only occasionally or sometimes
I feel humiliated or shamed by other people only occasionally or sometimes
I am able to do most things I need to often, most or all the time
I never make plans to end my life

Health state 5
I feel terribly alone and isolated only occasionally or sometimes
I feel panic or terror only occasionally or sometimes
I feel humiliated or shamed by other people only occasionally or sometimes
I am able to do most things I need to only occasionally or sometimes
I never make plans to end my life

Health state 6
I feel terribly alone and isolated often, most or all the time I feel panic or terror only occasionally or sometimes I feel humiliated or shamed by other people only occasionally or sometimes I am able to do most things I need to only occasionally or sometimes I never make plans to end my life

Health state 7
I feel terribly alone and isolated often, most or all the time I feel panic or terror often, most or all the time I feel humiliated or shamed by other people only occasionally or sometimes I am able to do most things I need to only occasionally or sometimes I never make plans to end my life

Health state 8
I feel terribly alone and isolated often, most or all the time I feel panic or terror often, most or all the time I feel humiliated or shamed by other people often, most or all the time I am able to do most things I need to only occasionally or sometimes I never make plans to end my life

Health state 9
I feel terribly alone and isolated often, most or all the time I feel panic or terror often, most or all the time I feel humiliated or shamed by other people often, most or all the time I am able to do most things I need to only occasionally or sometimes I make plans to end my life only occasionally or sometimes

Health state 10
I feel terribly alone and isolated often, most or all the time I feel panic or terror often, most or all the time I feel humiliated or shamed by other people often, most or all the time I am not able to do most things I need to I make plans to end my life only occasionally or sometimes

Health state 11
I feel terribly alone and isolated often, most or all the time I feel panic or terror often, most or all the time I feel humiliated or shamed by other people often, most or all the time I am not able to do most things I need to I make plans to end my life often, most or all the time

Appendix 13. SPSS syntax for calculation of CORE-6D utility values from CORE-OM data

```
*Deriving a CORE-6D utility score from CORE-OM data*

*Author: Ifigeneia Mavranezouli, December 2010*
*Based on Mavranezouli et al., MDM 2013; 33(3): 381-95*

*NOTES*
*Utility score is based on CORE-OM items 1, 8, 15, 16, 21, 33
*These are assumed to be named cof01, cos08, cos15, cor16, cof21, cof33 according to their domain
*Scores for item 21 are assumed to be reversed already [positive item]
*These must be converted to CORE-6D items co6D01, co6D08, co6D15, co6D16, co6D21, co6D33 (different levels of response)
*missing data are coded as 9
*total raw score of CORE-6D is named CORE6Dsc (note this EXCLUDES physical item 8)
*total utility score of CORE-6D is named CORE6Dut

*Converting CORE-OM items into CORE-6D items*

Compute co6D01=cof01.
IF (cof01=0) co6D01=0.
IF (cof01=1) co6D01=1.
IF (cof01=2) co6D01=1.
IF (cof01=3) co6D01=2.
IF (cof01=4) co6D01=2.
IF (cof01=9) co6D01=9.
IF (cof01<0) OR (cof01>4) co6D01=9.
Execute.

Compute co6D08=cos08.
IF (cos08=0) co6D08=0.
IF (cos08=1) co6D08=1.
IF (cos08=2) co6D08=1.
IF (cos08=3) co6D08=2.
IF (cos08=4) co6D08=2.
IF (cos08=9) co6D08=9.
IF (cos08<0) OR (cos08>4) co6D08=9.
Execute.
```

```
Compute co6D15=cos15.
IF (cos15=0) co6D15=0.
IF (cos15=1) co6D15=1.
IF (cos15=2) co6D15=1.
IF (cos15=3) co6D15=2.
IF (cos15=4) co6D15=2.
IF (cos15=9) co6D15=9.
IF (cos15<0) OR (cos15>4) co6D15=9.
Execute.

Compute co6D16=cor16.
IF (cor16=0) co6D16=0.
IF (cor16=1) co6D16=1.
IF (cor16=2) co6D16=1.
IF (cor16=3) co6D16=2.
IF (cor16=4) co6D16=2.
IF (cor16=9) co6D16=9.
IF (cor16<0) OR (cor16>4) co6D16=9.
Execute.

Compute co6D21=cof21.
IF (cof21=0) co6D21=0.
IF (cof21=1) co6D21=0.
IF (cof21=2) co6D21=1.
IF (cof21=3) co6D21=1.
IF (cof21=4) co6D21=2.
IF (cof21=9) co6D21=9.
IF (cof21<0) OR (cof21>4) co6D21=9.
Execute.

Compute co6D33=cof33.
IF (cof33=0) co6D33=0.
IF (cof33=1) co6D33=1.
IF (cof33=2) co6D33=1.
IF (cof33=3) co6D33=2.
IF (cof33=4) co6D33=2.
IF (cof33=9) co6D33=9.
IF (cof33<0) OR (cof33>4) co6D33=9.
Execute.
```

```
*Estimating total raw CORE-6D score – excluding physical item*

COMPUTE CORE6Dsc=(co6D01+co6D15+co6D16+co6D21+co6D33).
IF (co6D01=9) CORE6Dsc=99.
IF (co6D15=9) CORE6Dsc=99.
IF (co6D16=9) CORE6Dsc=99.
IF (co6D21=9) CORE6Dsc=99.
IF (co6D33=9) CORE6Dsc=99.
IF (co6D08=9) CORE6Dsc=99.
EXECUTE.

*Estimating total utility CORE-6D score*

Compute CORE6Dut=(CORE6Dsc+co6D08).
IF (CORE6Dsc=0) AND (co6D08=0) CORE6Dut=0.95.
IF (CORE6Dsc=1) AND (co6D08=0) CORE6Dut=0.94.
IF (CORE6Dsc=2) AND (co6D08=0) CORE6Dut=0.87.
IF (CORE6Dsc=3) AND (co6D08=0) CORE6Dut=0.80.
IF (CORE6Dsc=4) AND (co6D08=0) CORE6Dut=0.72.
IF (CORE6Dsc=5) AND (co6D08=0) CORE6Dut=0.64.
IF (CORE6Dsc=6) AND (co6D08=0) CORE6Dut=0.55.
IF (CORE6Dsc=7) AND (co6D08=0) CORE6Dut=0.47.
IF (CORE6Dsc=8) AND (co6D08=0) CORE6Dut=0.38.
IF (CORE6Dsc=9) AND (co6D08=0) CORE6Dut=0.30.
IF (CORE6Dsc=10) AND (co6D08=0) CORE6Dut=0.24.
IF (CORE6Dsc=0) AND (co6D08=1) CORE6Dut=0.92.
IF (CORE6Dsc=1) AND (co6D08=1) CORE6Dut=0.90.
IF (CORE6Dsc=2) AND (co6D08=1) CORE6Dut=0.84.
IF (CORE6Dsc=3) AND (co6D08=1) CORE6Dut=0.77.
IF (CORE6Dsc=4) AND (co6D08=1) CORE6Dut=0.69.
IF (CORE6Dsc=5) AND (co6D08=1) CORE6Dut=0.61.
IF (CORE6Dsc=6) AND (co6D08=1) CORE6Dut=0.52.
IF (CORE6Dsc=7) AND (co6D08=1) CORE6Dut=0.43.
IF (CORE6Dsc=8) AND (co6D08=1) CORE6Dut=0.35.
IF (CORE6Dsc=9) AND (co6D08=1) CORE6Dut=0.26.
IF (CORE6Dsc=10) AND (co6D08=1) CORE6Dut=0.20.
IF (CORE6Dsc=0) AND (co6D08=2) CORE6Dut=0.81.
IF (CORE6Dsc=1) AND (co6D08=2) CORE6Dut=0.80.
IF (CORE6Dsc=2) AND (co6D08=2) CORE6Dut=0.73.
IF (CORE6Dsc=3) AND (co6D08=2) CORE6Dut=0.66.
```

```
IF (CORE6Dsc=4) AND (co6D08=2) CORE6Dut=0.58.
IF (CORE6Dsc=5) AND (co6D08=2) CORE6Dut=0.50.
IF (CORE6Dsc=6) AND (co6D08=2) CORE6Dut=0.41.
IF (CORE6Dsc=7) AND (co6D08=2) CORE6Dut=0.32.
IF (CORE6Dsc=8) AND (co6D08=2) CORE6Dut=0.24.
IF (CORE6Dsc=9) AND (co6D08=2) CORE6Dut=0.16.
IF (CORE6Dsc=10) AND (co6D08=2) CORE6Dut=0.10.
EXECUTE.

VARIABLE LABELS CORE6Dut "CORE-6D preference-based utility".
EXECUTE.
```

Chapter 10. References

Ali S, Ronaldson S (2012) Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. British Medical Bulletin, 103:21-44.

Altman DG (1991) Practical statistics for medical research. London: Chapman & Hall.

Amin S, Singh SP, Croudace T, Jones P, Medley I, Harrison G (1999) Evaluating the Health of the Nation Outcome Scales. Reliability and validity in a three-year follow-up of first-onset psychosis. British Journal of Psychiatry, 174:399-403.

Andreas S, Harries-Hedder K, Schwenk W, Hausberg M, Koch U, Schulz H (2010) Is the Health of the Nation Outcome Scales appropriate for the assessment of symptom severity in patients with substance-related disorders? Journal of Substance Abuse Treatment, 39:32-40.

Andrich D (1978) A rating formulation for ordered response categories. Psychometrika, 43:561-573.

Andrich D, Lyne A, Sheridan B, Luo G (2003) RUMM2020. Perth: RUMM Laboratory Pty Ltd.

Audin K, Margison FR, Clark JM, Barkham M (2001) Value of HoNOS in assessing patient change in NHS psychotherapy and psychological treatment services. British Journal of Psychiatry, 178:561-566.

Baker FB (2001) The Basics of Item Response Theory (2nd edition). ERIC Clearinghouse on Assessment and Evaluation.

Bansback N, Brazier J, Tsuchiya A, Anis A (2012) Using a discrete choice experiment to estimate health state utility values. Journal of Health Economics, 31:306-318.

Barkham M, Bewick B, Mullin T, Gilbody S, Connell J, Cahill J, et al. (2013) The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. Counselling and Psychotherapy Research, 13:3-13.

Barkham M, Culverwell A, Spindler K, Twigg E (2005a) The CORE-OM in an older adult population: psychometric status, acceptability, and feasibility. Aging and Mental Health, 9:235-245.

Barkham M, Evans C, Margison F, McGrath G, Mellor-Clark J, Milne D, et al. (1998) The rationale for developing and implementing core batteries in service settings and psychotherapy outcome research. Journal of Mental Health, 7:35-47.

Barkham M, Gilbert N, Connell J, Marshall C, Twigg E (2005b) 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.

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, 69:184-196.

Barkham M, Mellor-Clark J, Connell J, Cahill J (2006) A core approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. Counselling and Psychotherapy Research, 6:3-15.

Bartlett MS (1950) Tests of significance in factor analysis. British Journal of Statistical Psychology, 3:77-85.

Bebbington P, Brugha T, Hill T, Marsden L, Window S (1999) Validation of the Health of the Nation Outcome Scales. British Journal of Psychiatry, 174:389-394.

Bennett KJ, Torrance GW, Boyle MH, Guscott R (2000a) Cost-utility analysis in depression: the McSad utility measure for depression health states. Psychiatric Services, 51:1171-1176.

Bennett KJ, Torrance GW, Boyle MH, Guscott R, Moran LA (2000b) Development and testing of a utility measure for major, unipolar depression (McSad). Quality of Life Research, 9:109-120.

Bick D, MacArthur C (1995) The extent, severity and effect of health problems after childbirth. British Journal of Midwifery, 3:27-31.

Bland JM, Altman DG (1995) Multiple significance tests: The Bonferroni method. BMJ, 310:170.

Bleichrodt H (2002) A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Economics, 11:447-456.

Bond TG (1994) Too many factors? Rasch Measurement Transactions, 8:347.

Bond TG, Fox CM (2007) Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Bowling A (1997) Measuring health: a review of quality of life measurement scales. Buckingham/Bristol: Open University Press.

Brazier J (2008) Measuring and valuing mental health for use in economic evaluation. Journal of Health Services Research and Policy, 13(Suppl 3):70-75.

Brazier J (2010) Is the EQ-5D fit for purpose in mental health? British Journal of Psychiatry, 197:348-349.

Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. (2005a) Should patients have a greater role in valuing health states? Applied Health Economics and Health Policy, 4:201-208.

Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, O'Cathain A, et al. (2014) Validating generic preference-based measures of health in mental health populations and estimating mapping functions for widely used specific measures. Health Technology Assessment, 18(34).

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, 28:113-126.

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).

Brazier J, Fitzpatrick R (2002) Measures of health-related quality of life in an imperfect world: a comment on Dowie. Health Economics, 11:17-19.

Brazier J, Roberts J, Deverill M (2002) The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics, 21:271-292.

Brazier J, Rowen D, Tsuchiya A, Yang Y, Young T (2011) The impact of adding an extra dimension to a preference-based measure. Social Science and Medicine, 73:245-253.

Brazier J, Tsuchiya A (2010) Preference-based condition-specific measures of health: what happens to cross programme comparability? Health Economics, 19:125-129.

Brazier J, Usherwood T, Harper R, Thomas K (1998) Deriving a preference-based single index from the UK SF-36 Health Survey. Journal of Clinical Epidemiology, 51:1115-1128.

Brazier JE, Dixon S, Ratcliffe J (2009) The role of patient preferences in cost-effectiveness analysis: a conflict of values? PharmacoEconomics, 27:705-712.

Brazier JE, Ratcliffe J, Salomon J, Tsuchiya A (2007) Measuring and valuing health benefits for economic evaluation. Oxford/New York: Oxford University Press.

Brazier JE, Roberts J (2004) The estimation of a preference-based measure of health from the SF-12. Medical Care, 42:851-859.

Brazier JE, Roberts J, Platts M, Zoellner YF (2005b) Estimating a preference-based index for a menopause specific health quality of life questionnaire. Health and Quality of Life Outcomes, 3:13.

Brazier JE, Rowen D (2011) NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. Sheffield: ScHARR, Decision Support Unit, University of Sheffield. Available from: http://www.nicedsu.org.uk. [Accessed 23 February 2013]

Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young TA, Yang Y, et al. (2012) Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome). Health Technology Assessment, 16(32).

Brazier JE, Yang Y, Tsuchiya A, Rowen DL (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, 11:215-225.

Briggs A, Wild D, Lees M, Reaney M, Dursun S, Parry D, et al. (2008) Impact of schizophrenia and schizophrenia treatment-related adverse events on quality of life: direct utility elicitation. Health and Quality of Life Outcomes, 6:105.

Brodersen J, Meads D, Kreiner S, Thorsen H, Doward L, McKenna S (2007) Methodological aspects of differential item functioning in the Rasch model. Journal of Medical Economics, 10:309-324.

Brooks M, Davies S, Twigg E (2013) A measure for feelings - using inclusive research to develop a tool for evaluating psychological therapy (Clinical Outcomes in Routine Evaluation - Learning Disability). British Journal of Learning Disabilities, 41:320-329.

Brooks R. EuroQol: the current state of play (1996) Health Policy, 37:53-72.

Brooks R (2000) The reliability and validity of the Health of the Nation Outcome Scales: validation in relation to patient derived measures. Australian and New Zealand Journal of Psychiatry, 34:504-511.

Brown S, Lumley J (1998) Maternal health after childbirth: results of an Australian population based survey. British Journal of Obstetrics and Gynaecology, 105:156-161.

Burns A, Beevor A, Lelliott P, Wing J, Blakey A, Orrell M, et al. (1999) Health of the Nation Outcome Scales for elderly people (HoNOS 65+). British Journal of Psychiatry, 174:424-427.

Burns T, Patrick D (2007) Social functioning as an outcome measure in schizophrenia studies. Acta Psychiatrica Scandinavica, 116:403-418.

Cahill J, Barkham M, Stiles WB, Twigg E, Rees A, Hardy GE, et al. (2006) Convergent validity of the CORE measures with measures of depression for clients in brief cognitive therapy for depression. Journal of Counseling Psychology, 53:253-259.

Cattell RB (1966) The scree test for the number of factors. Multivariate Behavioral Research, 1:245-276.

Cerny CA, Kaiser HF (1977) A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivariate Behavioral Research, 12:43-47.

Charlwood P, Mason A, Goldacre M, Cleary R, Wilkinson E (eds) (1999) Health Outcome Indicators: Severe mental illness. Report of a working group to the Department of Health. Oxford: National Centre for Health Outcomes Development.

Chatfield C, Collins AJ (1980). Introduction to multivariate analysis. Cambridge: Chapman and Hall, University Press.

Chisholm D, Healey A, Knapp M (1997) QALYs and mental health care. Social Psychiatry and Psychiatric Epidemiology, 32:68-75.

Cohen J (1988) Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Connell J, Barkham M, Stiles WB, Twigg E, Singleton N, Evans O, et al. (2007) Distribution of CORE-OM scores in a general population, clinical cut-off points and comparison with the CIS-R. British Journal of Psychiatry, 190:69-74.

Connell J, Brazier J, O'Cathain A, Lloyd-Jones M, Paisley S (2012) Quality of life of people with mental health problems: a synthesis of qualitative research. Health and Quality of Life Outcomes, 10:138.

Connell J, Grant S, Mullin T (2006) Client initiated termination of therapy at NHS primary care counselling services. Counselling and Psychotherapy Research, 6:60-67.

CORE System Group (1999) CORE system user manual. Leeds: Psychological Therapies Research Centre, University of Leeds.

Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR (1992) Quality-of-life assessment: can we keep it simple? Journal of the Royal Statistical Society Series A, 155:353-393.

Crawford M, Thana L, Patterson S, Weaver T, Robotham D, Dhillon K, et al. (2010) Outcome measurement in mental health: the views of service users.

Report submitted to Mental Health Research Network. London: NIHR Mental Health Research Network.

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika, 16:297-334.

de Bekker-Grob EW, Ryan M, Gerard K (2012) Discrete choice experiments in health economics: a review of the literature. Health Economics, 21:145-172.

DeCoster J (1998) Overview of Factor Analysis. Available from http://www.stat-help.com/factor.pdf [Accessed 30 March 2010]

Department of Health (1999) A National Service Framework for Mental Health. Modern Standards & Service Models. London: Department of Health.

Department of Health (2008) Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). London: Department of Health.

Department of Health, Joint Surveys Unit of Social and Community Planning Research, Department of Epidemiology and Public Health UCL (1998) Health Survey for England 1996. London: the Stationary Office.

Dickens G, Sugarman P, Walker L (2007) HoNOS-secure: a reliable outcome measure for users of secure and forensic mental health services. Journal of Forensic Psychiatry and Psychology, 18:507-514.

Dolan P (1997) Modeling valuations for EuroQol health states. Medical Care, 35:1095-1108.

Dolan P (2000) Effect of age on health state valuations. Journal of Health Services Research and Policy, 5:17-21.

Dolan P (2001) Output measures and valuation in health. In: Drummond M, McGuire A (eds). Economic evaluation in health care. Merging theory with practice, p 46-67. New York, NY: Oxford University Press.

Dolan P, Gudex C (1995) Time preference, duration and health state valuations. Health Economics, 4:289-299.

Dolan P, Gudex C, Kind P, Williams A (1996) The time trade-off method: results from a general population study. Health Economics, 5:141-154.

Dolan P, Roberts J (2002) To what extent can we explain time trade-off values from other information about respondents? Social Science and Medicine, 54:919-929.

Dowie J (2002a) 'Decision validity...': A rejoinder. Health Economics, 11:21-22.

Dowie J (2002b) Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions. Health Economics, 11:1-8.

Drummond MF, Schulpher MJ, Torrance GW, O'Brien BJ, Stoddard GL (2005) Methods for the economic evaluation of health care programmes. Oxford/New York: Oxford University Press.

Duke B (2010) HoNOS in the consultation liaison psychiatry setting: is it valid? Australasian Psychiatry, 18:547-550.

Eagar K, Trauer T, Mellsop G (2005) Performance of routine outcome measures in adult mental health care. Australian and New Zealand Journal of Psychiatry, 39:713-718.

Espallargues M, Czoski-Murray CJ, Bansback NJ, Carlton J, Lewis GM, Hughes LA, et al. (2005) The impact of age-related macular degeneration on health status utility values. Investigative Ophthalmology and Visual Science, 46:4016-4023.

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.

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 and Psychotherapy, 10:374-388.

Evans C, Mellor-Clark J, Margison F, Barkham M, McGrath G, Connell J, et al. (2000) Clinical Outcomes in Routine Evaluation: The CORE-OM. Journal of Mental Health, 9:247-255.

Feeny D (2002) Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions". Health Economics, 11:13-16.

Feeny D (2013) Standardization and regulatory guidelines may inhibit science and reduce the usefulness of analyses based on the application of preference-based measures for policy decisions. Medical Decision Making, 33:316-319.

Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. (2002) Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Medical Care, 40:113-128.

Ferreira LN, Ferreira PL, Pereira LN, Rowen D, Brazier JE (2013) Exploring the consistency of the SF-6D. Value in Health, 16:1023-1031.

Fisher WJ (1992) Reliability Statistics. Rasch Measurement Transactions, 6:238.

Fitzpatrick R, Bowling A, Gibbons E, Haywood K, Jenkinson C, Mackintosh A, et al. (2007) A structured review of patient-reported measures in relation to selected chronic conditions, perceptions of quality of care and carer impact. Oxford: National Centre for Health Outcomes Development.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR (1998) Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment, 2(14).

Fleminger S, Leigh E, Eames P, Langrell L, Nagraj R, Logsdail S (2005) HoNOS-ABI: a reliable measure of neuropsychiatric sequelae to brain injury? Psychiatric Bulletin, 29:53-55.

Fletcher AE (1988) Measurement of quality of life in clinical trials of therapy. Recent Results in Cancer Research, 111:216-230.

Flynn K (2002) Outcome measurement in VHA mental health services. An overview and series of diagnosis specific short reports. Part 1. Overview: Global measures of mental health status, psychiatric symptoms, and functioning. Boston, MA: VA Technology Assessment Program, Office of Patient Care Services.

Flynn K (2003) Outcome measurement in schizophrenia (Number 2 in a series: outcome measurement in VHA mental health services). Boston, MA: VA Technology Assessment Program, Office of Patient Care Services.

Flynn K (2004) Outcome measurement in major depression (Number 3 in a Series: Outcome measurement in VHA mental health services). Boston, MA: VA Technology Assessment Program, Office of Patient Care Services.

Fonagy P, Matthews R, Pilling S (2004) The Mental Health Outcomes Measurement Initiative: Report from the Chair of the Outcomes Reference Group. London: National Collaborating Centre for Mental Health.

Franks P, Lubetkin EI, Gold MR, Tancredi DJ, Jia H (2004) Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample. Medical Decision Making, 24:247-254.

Garau M, Shah KK, Mason AR, Wang Q, Towse A, Drummond MF (2011) Using QALYs in cancer: a review of the methodological limitations. PharmacoEconomics, 29:673-685.

Gilbody S, Richards D, Barkham M (2007) Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-OM. British Journal of General Practice, 57:650-652.

Gilbody SM, House AO, Sheldon TA (2003) Outcomes Measurement in Psychiatry. A critical review of outcomes measurement in psychiatric research and practice. York: University of York.

Gold MR, Patrick DL, Torrance GW, Fryback DG, Hadorn DC, Kamlet MS, et al. (1996) Identifying and valuing outcomes. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (eds). Cost-Effectiveness in Health and Medicine, p 82-123. Oxford: Oxford University Press.

Gowers SG, Harrington RC, Whitton A, Lelliott P, Beevor A, Wing J, et al. (1999) Brief scale for measuring the outcomes of emotional and behavioural disorders in children. Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA). British Journal of Psychiatry, 174:413-416.

Gray AM, Rivero-Arias O, Clarke PM (2006) Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. Medical Decision Making, 26:18-29.

Gray P, Mellor-Clark J (2007) CORE: A decade of development. Rugby: Penny Gray on behalf of CORE IMS.

Green C, Brazier J, Deverill M (2000) Valuing health-related quality of life. A review of health state valuation techniques. PharmacoEconomics, 17:151-165.

Gudex C (1994) Time Trade-Off User Manual: Props and Self-Completion Methods. York: Centre for Health Economics, University of York.

Gudex C, Dolan P, Kind P, Williams A (1996) Health state valuations from the general public using the visual analogue scale. Quality of Life Research, 5:521-531.

Guttman LA (1950) The basis for scalogram analysis. In: Stouffer SA, Guttman LA, Schuman EA (eds). Measurement and prediction. Volume 4 of Studies in social psychology in World War II, p 60-90. Princeton, NJ: Princeton University Press.

Guyatt G (2002) Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions". Health Economics, 11:9-12.

Guyatt GH, Feeny DH, Patrick DL (1993) Measuring health-related quality of life. Annals of Internal Medicine, 118:622-629.

Hampson M, Killaspy H, Mynors-Wallis L, Meier R (2011) Outcome measures recommended for use in adult psychiatry. London: Royal College of Psychiatrists.

Harper R, Brazier JE, Waterhouse JC, Walters SJ, Jones NM, Howard P (1997) Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. Thorax, 52:879-887.

Harrell FE Jr (2001) Regression Modelling Strategies: with applications to linear models, logistic regression, and survival analysis. New York (NY): Springer.

Harvey RJ, Hammer AL (1999) Item Response Theory. The Counseling Psychologist, 27:353-383.

Haywood KL, Garratt AM, Lall R, Smith JF, Lamb SE (2008) EuroQol EQ-5D and condition-specific measures of health outcome in women with urinary incontinence: reliability, validity and responsiveness. Quality of Life Research, 17:475-483.

Hedayat AS, Sloane NJA, Stufken J (1999) Orthogonal arrays, theory and applications. New York, NY: Springer.

Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. (2011) Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research, 20:1727-1736.

HM Government, Department of Health (2011) No Health Without Mental Health: a cross-government mental health outcomes strategy for people of all ages. London: Department of Health.

Horn JL (1965) A rationale and test for the number of factors in factor analysis. Psychometrika, 30:179-185.

Hornberger JC, Redelmeier DA, Petersen J (1992) Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. Journal of Clinical Epidemiology, 45:505-512.

Hounsome N, Orrell M, Edwards RT (2011) EQ-5D as a quality of life measure in people with dementia and their carers: Evidence and key issues. Value in Health, 14:390-399.

Howard KI, Lueger RJ, Maling MS, Martinovich Z (1993) A phase model of psychotherapy outcome: causal mediation of change. Journal of Consulting and Clinical Psychology, 61:678-685.

IAPT (2008) Improving Access to Psychological Therapies (IAPT) Outcomes Toolkit 2008/9. London: National IAPT Programme Team.

IAPT (2011) The IAPT data handbook. Guidance on recording and monitoring outcomes to support local evidence-based practice. Version 2.0.1. London: National IAPT Programme Team.

IBM Corp (2010) IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp.

Jacobs R (2009) Investigating Patient Outcome Measures in Mental Health. CHE Research Paper Number 48. York: Centre for Health Economics, University of York.

Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Controlled Clinical Trials, 10:407-415.

Janson-Bjerklie S, Ferketich S, Benner P, Becker G (1992) Clinical markers of asthma severity and risk: importance of subjective as well as objective factors. Heart and Lung, 21:265-272.

Jenkinson C, Coulter A, Reeves R, Bruster S, Richards N (2003) Properties of the Picker Patient Experience questionnaire in a randomized controlled trial of long versus short form survey instruments. Journal of Public Health Medicine, 25:197-201.

Jenkinson C, McGee H (1998) Health Status Measurement. A brief but critical introduction. Abington: Radcliffe Medical Press Ltd.

Jones SH, Thornicroft G, Coffey M, Dunn G (1995) A brief mental health outcome scale-reliability and validity of the Global Assessment of Functioning (GAF). British Journal of Psychiatry, 166:654-659.

Kaiser HF (1960) The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20:141-151.

Kharroubi SA, Brazier JE, Roberts J, O'Hagan A (2007) Modelling SF-6D health state preference data using a nonparametric Bayesian method. Journal of Health Economics, 26:597-612.

Kieffer KM (1998) Orthogonal versus Oblique Factor: a Review of the Literature regarding the Pros and Cons. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA. Available from: http://files.eric.ed.gov/fulltext/ED427031.pdf [Accessed 4 February 2009]

Kind P, Hardman G, Macran S (1999) UK population norms for EQ-5D. York: Centre for Health Economics Discussion Paper Series, University of York.

Kind P, Macran S (2005) Eliciting social preference weights for functional assessment of cancer therapy-lung health states. PharmacoEconomics, 23:1143-1153.

Kirshner B, Guyatt G (1985) A methodological framework for assessing health indices. Journal of Chronic Diseases, 38:27-36.

Kisely S, Campbell LA, Cartwright J, Cox M, Campbell J (2010) Do the Health of the Nation Outcome Scales measure outcome? Canadian Journal of Psychiatry, 55:431-439.

Kisely S, Campbell LA, Crossman D, Gleich S, Campbell J (2007) Are the Health of the Nation Outcome Scales a valid and practical instrument to measure outcomes in North America? A three-site evaluation across Nova Scotia. Community Mental Health Journal, 43:91-107.

Knapp M, Mangalore R (2007) "The trouble with QALYs..." Epidemiologia e Psichiatria Sociale, 16:289-293.

Kobelt G, Kirchberger I, Malone-Lee J (1999) Review. Quality-of-life aspects of the overactive bladder and the effect of treatment with tolterodine. BJU International, 83:583-590.

Kodagalli A, Mynors-Wallis L, Cope D, Ogollah R, Immins T (2012) Patient-reported outcome measures versus clinician-measured outcomes in community psychiatric practice. The Psychiatrist Online, 36:61-64.

Kok ET, McDonnell J, Stolk EA, Stoevelaar HJ, Busschbach JJ (2002) The valuation of the International Prostate Symptom Score (IPSS) for use in economic evaluations. European Urology, 42:491-497.

Kowalski JW, Rentz AM, Walt JG, Lloyd A, Lee J, Young TA, et al. (2012) Rasch analysis in the development of a simplified version of the national eye institute visual-function questionnaire-25 for utility estimation. Quality of Life Research, 21:323-334.

Krabbe PF, Stouthard ME, Essink-Bot ML, Bonsel GJ (1999) The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. Journal of Clinical Epidemiology, 52:293-301.

Lamers LM, Uyl-de Groot CA, Buijt I (2007) The use of disease-specific outcome measures in cost-utility analysis: The development of Dutch societal preference weights for the FACT-L scale. PharmacoEconomics, 25:591-603.

Leach C, Lucock M, Barkham M, Noble R, Clarke L, Iveson S (2005) Assessing risk and emotional disturbance using the CORE-OM and HoNOS outcome measures at the interface between primary and secondary mental healthcare. Psychiatric Bulletin, 29:419-422.

Leach C, Lucock M, Barkham M, Stiles WB, Noble R, Iveson S (2006) Transforming between Beck Depression Inventory and CORE-OM scores in routine clinical practice. British Journal of Clinical Psychology, 45:153-166.

Lee TT, Ziegler JK, Sommi R, Sugar C, Mahmoud R, Lenert LA (2000) Comparison of preferences for health outcomes in schizophrenia among stakeholder groups. Journal of Psychiatric Research, 34:201-210.

Lehman AF, Lasalvia A (2010) Measures of quality of life for people with severe mental disorders. In: Tansella M, Thornicroft G (eds). Mental Health Outcome Measures (3rd edition), p. 135-168. London: Gaskell.

Leidy NK, Revicki DA, Geneste B (1999) Recommendations for evaluating the validity of quality of life claims for labeling and promotion. Value in Health, 2:113-127.

Lenert LA, Sherbourne CD, Sugar C, Wells KB (2000a) Estimation of utilities for the effects of depression from the SF-12. Medical Care, 38:763-770.

Lenert LA, Sturley AP, Rapaport MH, Chavez S, Mohr PE, Rupnow M (2004) Public preferences for health states with schizophrenia and a mapping function to estimate utilities from positive and negative symptom scale scores. Schizophrenia Research, 71:155-165.

Lenert LA, Treadwell JR, Schwartz CE (1999) Associations between health status and utilities: implications for policy. Medical Care, 37:479-489.

Lenert LA, Ziegler J, Lee T, Sommi R, Mahmoud R (2000b) Differences in health values among patients, family members, and providers for outcomes in schizophrenia. Medical Care, 38:1011-1021.

Lewis G, Pelosi AJ, Araya R, Dunn G (1992) Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. Psychological Medicine, 22:465-486.

Lin MR, Yu WY, Wang SC (2012) Examination of assumptions in using time tradeoff and standard gamble utilities in individuals with spinal cord injury. Archives of Physical Medicine and Rehabilitation, 93:245-252.

Linacre JM (2003) Rasch Power Analysis: Size vs. Significance: Standardized Chi-Square Fit Statistic. Rasch Measurement Transactions, 17:918.

Long JS (1997) Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications.

Luo G (2005) The relationship between the Rating Scale and Partial Credit Models and the implication of disordered thresholds of the Rasch models for polytomous responses. Journal of Applied Measurement, 6:443-455.

Marshall K, Willoughby-Booth S (2007) Modifying the Clinical Outcomes in Routine Evaluation measure for use with people who have a learning disability. British Journal of Learning Disabilities, 35:107-112.

Mavranezouli I, Brazier JE, Rowen D, Barkham M (2012) Estimating a Preference-Based Index from the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM): Valuation of CORE-6D. Medical Decision Making, 33:381-395.

Mavranezouli I, Brazier JE, Young TA, Barkham M (2011) 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). Quality of Life Research, 20:321-333.

McCabe R, Saidi M, Priebe S (2007) Patient-reported outcomes in schizophrenia. British Journal of Psychiatry, 191:s21-s28.

McClelland R, Trimble P, Fox ML, Stevenson MR, Bell B (2000) Validation of an outcome scale for use in adult psychiatric practice. Quality in Health Care, 9:98-105.

McCrone P, Dhanasiri S, Patel A, Knapp M, Lawton-Smith S (2008) Paying the price. The cost of mental health care in England to 2026. London: The King's Fund, 2008.

McCrone P, Marks IM, Mataix-Cols D, Kenwright M, McDonough M (2009) Computer-aided self-exposure therapy for phobia/panic disorder: a pilot economic evaluation. Cognitive Behaviour Therapy, 38:91-99.

McDowell I, Newell C (1996) Measuring health: a guide to rating scales and questionnaires. Oxford/New York: Oxford University Press.

McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1:30-46.

McKenna SP, Ratcliffe J, Meads DM, Brazier JE (2008) Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. Health and Quality of Life Outcomes, 6:65.

McManus S, Meltzer H, Brugha T, Bebbington P, Jenkins R (2009) Adult psychiatric morbidity in England, 2007: Results of a household survey. Leeds: The NHS Information Centre for Health and Social Care.

McTaggart-Cowan H (2011) Elicitation of informed general population health state utility values: a review of the literature. Value in Health, 14:1153-1157.

McTaggart-Cowan H, Tsuchiya A, O'Cathain A, Brazier J (2011) Understanding the effect of disease adaptation information on general population values for hypothetical health states. Social Science and Medicine, 72:1904-1912.

McTaggart-Cowan HM, Brazier JE, Tsuchiya A (2010) Clustering Rasch Results: A Novel Method for Developing Rheumatoid Arthritis States for Use in Valuation Studies. Value in Health, 13:787-795.

McTaggart-Cowan HM, O'Cathain A, Tsuchiya A, Brazier JE (2012) Using mixed methods research to explore the effect of an adaptation exercise on general population valuations of health states. Quality of Life Research, 21:465-473.

Medical Research Council (2010) Review of mental health research: report of the Strategic Review Group 2010. London: Medical Research Council.

Mehrez A, Gafni A (1993) Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. Medical Decision Making, 13:287-292.

Mellor-Clark J, Jenkins AC, Evans R, Mothersole G, McInnes B (2006) Resourcing a CORE Network to develop a National Research Database to help enhance psychological therapy and counselling service provision. Counselling and Psychotherapy Research, 6:16-22.

Miller GA (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. Psychological Review, 63:81-97.

Mohr PE, Cheng CM, Claxton K, Conley RR, Feldman JJ, Hargreaves WA, et al. (2004) The heterogeneity of schizophrenia in disease states. Schizophrenia Research, 71:83-95.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research, 19:539-549.

Morrell CJ, Slade P, Warner R, Paley G, Dixon S, Walters SJ, et al. (2009a) Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. BMJ, 338:a3045.

Morrell CJ, Warner R, Slade P, Dixon S, Walters S, Paley G, et al. (2009b) Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation. The PoNDER trial. Health Technology Assessment, 13(30).

Mulhern B, Rowen D, Jacoby A, Marson T, Snape D, Hughes D, et al. (2012a) The development of a QALY measure for epilepsy: NEWQOL-6D. Epilepsy and Behavior, 24:36-43.

Mulhern B, Smith SC, Rowen D, Brazier JE, Knapp M, Lamping DL, et al. (2012b) Improving the measurement of QALYs in dementia: developing patient- and carer-reported health state classification systems using Rasch analysis. Value in Health, 15:323-333.

Mullin T, Barkham M, Mothersole G, Bewick BM, Kinder A (2006) Recovery and improvement benchmarks in routine primary care mental health settings. Counselling and Psychotherapy Research, 6:68-80.

MVH Group (1995) The measurement and valuation of health: Final report on the modelling of valuation tariffs. York: Centre for Health Economics, University of York.

NCCMH (2010a) Depression: the Treatment and Management of Depression in Adults. Updated edition. Leicester & London: The British Psychological Society and the Royal College of Psychiatrists.

NCCMH (2010b) Schizophrenia: Core Interventions in the Treatment and Management of Schizophrenia in Adults in Primary and Secondary Care. Updated edition. Leicester & London: The British Psychological Society and the Royal College of Psychiatrists.

NCCMH (2011) Generalised Anxiety Disorder in Adults: Management in Primary, Secondary and Community Care. Leicester & London: The British Psychological Society and the Royal College of Psychiatrists.

National Collaborating Centre for Mental Health (2013) Social anxiety disorder: recognition, assessment and treatment of social anxiety disorder. Leicester & London: The British Psychological Society and the Royal College of Psychiatrists.

National Collaborating Centre for Mental Health (2014) Bipolar disorder: The assessment and management of bipolar disorder in adults, children and young people, in primary and secondary care. Leicester & London: The British Psychological Society and the Royal College of Psychiatrists.

National Institute for Health and Care Excellence (2013) Guide to the Methods of Technology Appraisal 2013. London: National Institute for Health and Care Excellence.

National Institute for Health and Clinical Excellence (2008) Guide to the Methods of Technology Appraisal 2008. London: National Institute for Health and Clinical Excellence.

National Institute for Mental Health in England (2008) Mental Health Outcomes Compendium. London: Department of Health.

Neumann PJ, Goldie SJ, Weinstein MC (2000) Preference-based measures in economic evaluation in health care. Annual Review of Public Health, 21:587-611.

Oiesvold T, Bakkejord T, Sexton JA (2011) Concurrent validity of the Health of the Nation Outcome Scales compared with a patient-derived measure, the Symptom Checklist-90-Revised in out-patient clinics. Psychiatry Research, 187:297-300.

Orrell M, Yard P, Handysides J, Schapira R (1999) Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community. British Journal of Psychiatry, 174:409-412.

Overall JE, Gorham DR (1962) The Brief Psychiatric Rating Scale. Psychological Reports, 10:799-812.

Oxford Economics (2007) Mental Health and the UK economy. Oxford: Oxford Economics.

Page AC, Hooke GR, Rutherford EM (2001) Measuring mental health outcomes in a private psychiatric clinic: Health of the Nation Outcome Scales and Medical Outcomes Short Form SF-36. Australian and New Zealand Journal of Psychiatry, 35:377-381.

Pallant JF, Tennant A (2007) An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology, 46:1-18.

Papaioannou D, Brazier J, Parry G (2011) How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value in Health, 14:907-920.

Papaioannou D, Brazier J, Parry G (2013) How to Measure Quality of Life for Cost-Effectiveness Analyses of Personality Disorders: A Systematic Review. Journal of Personality Disorders, 27:383-401.

Parkin D, Devlin N (2006) Is there a case for using visual analogue scale valuations in cost-utility analysis? Health Economics, 15:653-664.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA (1994) Measuring preferences for health states worse than death. Medical Decision Making, 14:9-18.

Peasgood T, Brazier J, Papaioannou D (2012) A systematic review of the validity and responsiveness of EQ-5D and SF-6D for depression and anxiety. HEDS discussion paper 12/15. Sheffield: ScHARR, University of Sheffield.

Phelan M, Slade M, Thornicroft G, Dunn G, Holloway F, Wykes T, et al. (1995) The Camberwell Assessment of Need: the validity and reliability of an instrument to assess the needs of people with severe mental illness. British Journal of Psychiatry, 167:589-595.

Pirkis JE, Burgess PM, Kirk PK, Dodson S, Coombs TJ, Williamson MK (2005) A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. Health and Quality of Life Outcomes, 3:76.

Proctor G, Hargate R (2013). Quantitative and qualitative analysis of a set of goal attainment forms in primary care mental health services. Counselling and Psychotherapy Research, 13:235-241.

Quality and Outcomes Sub Group of the Product Review Group for Mental Health Payment by Results (2011) Payment by Results Quality and Outcomes Indicators. Leeds: Payment by results team, Department of Health.

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press.

Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M (2009) Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Economics, 18:1261-1276.

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC (1984) Preferences for health outcomes. Comparison of assessment methods. Medical Decision Making, 4:315-329.

Rentz AM, Kowalski JW, Walt JG,Hays RD, Brazier JE, Yu R, Lee P, Bressler N, Revicki DA (2014) Development of a preference-based index from the National Eye Institute Visual Function Questionnaire–25. JAMA Ophthalmology, 132(3):310-8.

Revicki DA, Hanlon J, Martin S, Gyulai L, Nassir GS, Lynch F, et al. (2005) Patient-based utilities for bipolar disorder-related health states. Journal of Affective Disorders, 87:203-210.

Revicki DA, Leidy NK, Brennan-Diemer F, Sorensen S, Togias A (1998a) Integrating patient preferences into health outcomes assessment: the multiattribute Asthma Symptom Utility Index. Chest, 114:998-1007.

Revicki DA, Leidy NK, Brennan-Diemer F, Thompson C, Togias A (1998b) Development and preliminary validation of the multiattribute Rhinitis Symptom Utility Index. Quality of Life Research, 7:693-702.

Revicki DA, Shakespeare A, Kind P (1996) Preferences for schizophrenia-related health states: a comparison of patients, caregivers and psychiatrists. International Clinical Psychopharmacology, 11:101-108.

Revicki DA, Wood M (1998) Patient-assigned health state utilities for depression-related outcomes: differences by depression severity and antidepressant medications. Journal of Affective Disorders, 48:25-36.

Richards A, Barkham M, Cahill J, Richards D, Williams C, Heywood P (2003) PHASE: a randomised, controlled trial of supervised self-help cognitive behavioural therapy in primary care. British Journal of General Practice, 53:764-770.

Robinson A, Dolan P, Williams A (1997) Valuing health status using VAS and TTO: what lies behind the numbers? Social Science and Medicine, 45:1289-1297.

Rowen D, Brazier J (2011) Health Utility Measurement. In: Glied S, Smith PC (eds). The Oxford Handbook of Health Economics, p. 788-813. Oxford: Oxford University Press.

Rowen D, Brazier J, Young T, Gaugris S, Craig BM, King MT, et al. (2011) Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Value in Health, 14:721-731.

Rowen D, Mulhern B, Banerjee S, Hout B, Young TA, Knapp M, et al. (2012) Estimating preference-based single index measures for dementia using DEMQOL and DEMQOL-Proxy. Value in Health, 15:346-356.

Roy A, Matthews H, Clifford P, Fowler V, Martin DM (2002) Health of the Nation Outcome Scales for People with Learning Disabilities (HoNOS-LD). British Journal of Psychiatry, 180:61-66.

Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC (1996) The role of cost-effectiveness analysis in health and medicine. Panel on Cost-Effectiveness in Health and Medicine. JAMA, 276:1172-1177.

Schaffer A, Levitt AJ, Hershkop SK, Oh P, MacDonald C, Lanctot K (2002) Utility scores of symptom profiles in major depression. Psychiatry Research, 110:189-197.

Sculpher M (2013) Methods development for health technology assessment: is it time to set priorities? Medical Decision Making, 33:313-315.

Sharma VK, Wilkinson G, Fear S (1999) Health of the Nation Outcome Scales: a case study in general psychiatry. British Journal of Psychiatry, 174:395-398.

Shumway M, Sentell T, Chouljian T, Tellier J, Rozewicz F, Okun M (2003) Assessing preferences for schizophrenia outcomes: comprehension and decision strategies in three assessment methods. Mental Health Services Research, 5:121-35.

Sinclair A, Barkham M, Evans C, Connell J, Audin K (2005) Rationale and development of a general population well-being measure: Psychometric status of the GP-CORE in a student sample. British Journal of Guidance and Councelling, 33:153-173.

Singleton N, Bumpstead R, O'Brien M, Lee A, Meltzer H (2001) Psychiatric morbidity among adults living in private households, 2000. London: The Stationary Office.

Singleton N, Lewis G (2003) Better or worse: a longitudinal study of the mental health of adults living in private households in Great Britain. London: The Stationary Office.

Slade M, Beck A, Bindman J, Thornicroft G, Wright S (1999) Routine clinical outcome measures for patients with severe mental illness: CANSAS and HoNOS. British Journal of Psychiatry, 174:404-408.

Slade M, Powell R, Rosen A, Strathdee G (2000) Threshold Assessment Grid (TAG): the development of a valid and brief scale to assess the severity of mental illness. Social Psychiatry and Psychiatric Epidemiology, 35:78-85.

Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M (2008) Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology, 8:33.

Smith EVJ (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement, 3:205-231.

Stata Corp (2007) Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Stevens K, Palfreyman S (2012) The use of qualitative methods in developing the descriptive systems of preference-based measures of health-related quality of life for use in economic evaluation. Value in Health, 15:991-998.

Stevens KJ, Brazier JE, McKenna SP, Doward LC, Cork MJ (2005) The development of a preference-based measure of health in children with atopic dermatitis. British Journal of Dermatology, 153:372-377.

Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Habbema JD, De Haes JC (1995) The "utility" of the Time Trade-Off method in cancer patients: feasibility and proportional Trade-Off. Journal of Clinical Epidemiology, 48:1207-1214.

Stiles WB, Barkham M, Mellor-Clark J, Connell J (2008) Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: replication in a larger sample. Psychological Medicine, 38:677-688.

Streiner DL, Norman GR (1995) Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press.

Sugar CA, Sturm R, Lee TT, Sherbourne CD, Olshen RA, Wells KB, et al. (1998) Empirically defined health states for depression from the SF-12. Health Services Research, 33:911-928.

Sundaram M, Smith MJ, Revicki DA, Elswick B, Miller LA (2009) Rasch analysis informed the development of a classification system for a diabetes-specific preference-based measure of health. Journal of Clinical Epidemiology, 62:845-856.

Sundaram M, Smith MJ, Revicki DA, Miller LA, Madhavan S, Hobbs G (2010) Estimation of a valuation function for a diabetes mellitus-specific preference-based measure of health: the Diabetes Utility Index. PharmacoEconomics, 28:201-216.

Tabachnick BG, Fidell LS (1996) Using multivariate statistics. New York, NY: Harper Collins.

Tennant A, Conaghan PG (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 and Rheumatism, 57:1358-1362.

Tennant A, McKenna SP, Hagell P (2004) Application of Rasch analysis in the development and application of quality of life instruments. Value in Health 2004, 7(Suppl 1): S22-S26.

Tennant A, Pallant JF (2006) Unidimensionality matters! (A tale of two smiths?). Rasch Measurement Transactions, 20:1048-1051.

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60:34-42.

Tesio L (2003) Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine, 35:105-115.

Testa MA, Anderson RB, Nackley JF, Hollenberg NK (1993) Quality of life and antihypertensive therapy in men. A comparison of captopril with enalapril. The Quality-of-Life Hypertension Study Group. New England Journal of Medicine, 328:907-913.

Thompson JF, Roberts CL, Currie M, Ellwood DA (2002) Prevalence and persistence of health problems after childbirth: associations with parity and method of birth. Birth, 29:83-94.

Tilling C, Devlin N, Tsuchiya A, Buckingham K (2010) Protocols for time tradeoff valuations of health states worse than dead: a literature review. Medical Decision Making, 30:610-619.

Tobin J (1958) Estimation of relationships for limited dependent variables. Econometrica, 26:24-36.

Torrance G (1984) Health states worse than death. In: van Eimeren W, Engelbrecht R, Flagle CD (eds). Third International Conference on System Science in Health Care, p. 1085-1089. Berlin: Springer.

Torrance GW (1976) Social preferences for health states: an empirical evaluation of three measurement techniques. Socio-economic Planning Sciences, 10:129-136.

Torrance GW (1986) Measurement of health state utilities for economic appraisal. Journal of Health Economics, 5:1-30.

Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q (1996) Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Medical Care, 34:702-722.

Torrance GW, Furlong W, Feeny D, Boyle M (1995) Multi-attribute preference functions. Health Utilities Index. PharmacoEconomics, 7:503-520.

Torrance GW, Thomas WH, Sackett DL (1972) A utility maximization model for evaluation of health care programs. Health Services Research, 7:118-133.

Trauer T, Callaly T, Hantz P, Little J, Shields R, Smith J (1999) Health of the Nation Outcome Scales. Results of the Victorian field trial. British Journal of Psychiatry, 174:380-388.

Tsuchiya A, Brazier J, McColl E, Parkin D (2002) Deriving preference-based condition-specific instruments: converting AQLQ into EQ-5D indices. HEDS Discussion Paper 02/01. Sheffield: ScHARR, University of Sheffield.

Twigg E, Barkham M, Bewick BM, Mulherm B, Connell J, Cooper M (2009) The Young Person's CORE: Development of a brief outcome measure for young people. Counselling and Psychotherapy Research, 9:160-168.

US Department of Health and Human Services, Food and Drug Administration (2009) Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Rockville, MD: US Department of Health and Human Services, Food and Drug Administration (FDA).

Ubel PA, Loewenstein G, Jepson C (2003) Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Quality of Life Research, 12:599-607.

van de Willige G, Wiersma D, Nienhuis FJ, Jenner JA (2005) Changes in quality of life in chronic psychiatric patients: a comparison between EuroQol (EQ-5D) and WHOQoL. Quality of Life Research, 14:441-451.

Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA (2012) Condition-specific preference-based measures: benefit or burden? Value in Health, 15:504-513.

von Neumann J, Morgenstern O (1953) Theory of games and economic behaviour. New York, NY: Wiley.

Walters SJ, Morrell CJ, Dixon S (1999) Measuring health-related quality of life in patients with venous leg ulcers. Quality of Life Research, 8:327-336.

Ware JE, Kolinski M, Keller SD (1995) How to score the SF-12 physical and mental health summaries: a user's manual. Boston, MA: The Health Institute, New England Medical Centre.

Ware JE, Sherbourne CD, Davies AR (1992) Developing and testing the MOS 20-Item Short-Form Health Survey: A general population application. In: Stewart AL, Ware JE (eds). Measuring functioning and well-being: the Medical Outcomes Study approach, p. 277-290. Durham, NC: Duke University Press.

Ware JE, Snow KK, Kolinski M, Gandek B (1993) SF-36 Health survey manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Centre.

Watkins M (2008) Monte Carlo PCA for Parallel Analysis 2.3. Available from: http://www.softpedia.com/progDownload/Monte-Carlo-PCA-for-Parallel-Analysis-Download-56312.html [Accessed 24 January 2010].

Weiss DJ, Yoes ME (1991) Item Response Theory. In: Hambleton RK, Zaal JN (eds). Advances in educational and psychological testing: Theory and applications, p. 69-95. New York, NY: Kluwer Academic/Plenum Publishers.

Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C (2003) Comparative responsiveness of generic and specific quality-of-life instruments. Journal of Clinical Epidemiology, 56:52-60.

Wing JK, Beevor AS, Curtis RH, Park SB, Hadden S, Burns A (1998) Health of the Nation Outcome Scales (HoNOS). Research and development. British Journal of Psychiatry, 172:11-18.

World Health Organization (1958) The first ten years of the World Health Organization. Geneva: World Health Organization.

World Health Organization (2008) The global burden of disease: 2004 update. Geneva: WHO Press.

World Health Organization (2010) Mental health: strengthening our response. Fact sheet No 220. Geneva: WHO.

World Health Organization (2013) Mental Health Action Plan 2013-2020. Geneva: WHO Document Production Services.

Wright BD (1996) Local dependency, correlations and principal components. Rasch Measurement Transactions, 10:509-511.

Yang Y, Brazier J, Tsuchiya A (2013a) Effect of Adding a Sleep Dimension to the EQ-5D Descriptive System: A "Bolt-On" Experiment. Medical Decision Making, 34:42-53.

Yang Y, Brazier J, Tsuchiya A, Coyne K (2009) Estimating a preference-based single index from the Overactive Bladder Questionnaire. Value in Health, 12:159-166.

Yang Y, Brazier JE, Tsuchiya A, Young TA (2011) Estimating a Preference-Based Index for a 5-Dimensional Health State Classification for Asthma Derived From the Asthma Quality of Life Questionnaire. Medical Decision Making, 31:281-291.

Yang Y, Longworth L, Brazier J (2013b) An assessment of validity and responsiveness of generic measures of health-related quality of life in hearing impairment. Quality of Life Research, 22:2813-2828.

Young T, Yang Y, Brazier JE, 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, 18:253-265.

Young TA, Rowen D, Norquist J, Brazier JE (2010) Developing preference-based health measures: using Rasch analysis to generate health state values. Quality of Life Research, 19:907-917.

Young TA, Yang Y, Brazier JE, Tsuchiya A (2011) 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, 31:195-210.

Zwick WR, Velicer WF (1986) Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99:432-442.