An investigation of methods for identifying and selecting bolt-on dimensions: the EQ-5D-5L case study

By:

Aureliano Paolo Finch

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Date of submission: October 2017

The University of Sheffield

Faculty of Medicine, Dentistry and Health

School of Health and Related Research

Contents

List of abbreviations

AQoL	Assessment of Quality of Life
CBA	Cost Benefit Analysis
CEA	Cost Effectiveness Analysis
CFA	Confirmatory Factor Analysis
CUA	Cost Utility Analysis
EFA	Exploratory Factor Analysis
GPBM	Generic Preference Based Measure
HRQoL	Health Related Quality of Life
HTA	Health Technology Assessment
HUI	Health Utility Index
ICECAP	ICEpop CAPability measure
MIC	Multi Instrument Comparison
NICE	National Institute for Health and Care Excellence
ONS	Office for National Statistics
PBM	Preference Based Measures
PCA	Principal Component Analysis
PWI	Personal well-being index
QALY	Quality Adjusted Life Years
SEM	Structural Equation Modeling
SG	Standard Gamble
SWB	Subjective Well-being
TTO	Time Trade Off
SWLS	Satisfaction with Life Scale
VAS	Visual Analogue Scale

Declaration

I declare that this thesis is my original work. None of the parts of this thesis has ever been submitted for a degree at this or another institution. All sources of assistance have been acknowledged. External sources of information are cited as appropriate.

My supervisors, Prof. John E Brazier and Dr. Clara Mukuria, provided guidance throughout the doctoral studies. They contributed to the design and interpretation of the empirical studies reported in Chapter 3, 4, 5 and 6, and should be considered secondary co-authors of those chapters. I held primary responsibility for all aspects of the research in this thesis, and should be considered the lead author of Chapters 3, 4, 5 and 6 and the sole authors of Chapters 1, 2 and 7. Prof. Jakob Bue Bjorner provided substantial suggestions in the interpretation and presentation of Chapter 4.

Acknowledgements

The University of Sheffield - School of Health and Related Research provided financial support for this research. Additional financial support for the study reported in Chapter 6 was provided through a grant awarded by the EuroQol Research Foundation. I would like to thank these institutions as they have been fundamental for conducting this program of study. The views expressed in this thesis, and any errors or omissions, are of the author only and do not necessarily reflect the positions of the funders.

There are numerous people that I have to thank and without whom I would have not been able to complete this work. First of all, I would like to express my foremost gratitude to my supervisors John E Brazier and Clara Mukuria. Their insightful suggestions, support and encouragement have resulted in the most rewarding professional experience of my life.

Apart from my supervisors, I would also like to thank Tessa Peasgood and Tracey Young for the incredibly valuable suggestions provided during my upgrade viva; Praveen Thokala for the great effort in reading the whole PhD thesis prior to its submission; Jakob Bue Bjorner for his indications on how to improve the factor analysis paper based on Chapter 5; Donna Rowen and Jill Carlton for their substantive suggestions made in the development of bolt-ons wording and in the design of the pairwise experiment of Chapter 6; Claire Beecroft and Diana Papaioannou for their help in developing the searching strategy of the overview of systematic reviews reported in Chapter 3; the administrative staff of the School of Health and Related Research for supporting the activities related to this study; and the rest of the faculty of the School of Health and Related Research for creating a pleasant and stimulating environment of work.

I would like to thank the anonymous reviewers who made helpful comments on manuscripts submitted to journals based on this research program and on the grant application. I also thank all of the respondents who took part in the empirical studies included in this research.

I want to acknowledge all the friends and PhD students that have shared this roller-coaster trip with me, and more particularly Peter, Milad, Beckie, Rach, Sam, Tom, Sharath, Kosuke, Glauco, Jacopo, Matteo, Serena, Erika, Abi, Martin, Alex and Fabrizio.

This thesis would not have been possible without my family. I would like to thank my mum and grandma for their constant encouragement and love that gave me the tranquillity to achieve my objectives; my dad for his love and for giving me the perseverance and strength to do what I needed; and Ale and Simo, as even away they are always brothers to me.

I need to give my biggest recognition to Hannah. The incredible time together, the laughs, the support, the insightful discussions and her affection made everything, even in the worst moments, pleasant and happy.

Publications

Finch AP, Brazier JE, Mukuria C, Bjorner J. An exploratory study on using principal component analysis and confirmatory factor analysis to identify bolt-on dimensions: The EQ-5D case study. *Value in Health*. 2017. *(in press).*

Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of Generic Preference-Based Measures? A systematic overview of reviews. *European Journal of Health Economics*. May 2017. *(in press).*

Finch AP, Brazier JE, Mukuria C. Selection of bolt-ons after factor analysis: are linear regression models a useful technique?. *Health and Quality of Life Outcomes*. 2017. 15; 185: A17.

Finch AP, Brazier JE, Mukuria CW. Validity and Responsiveness of Preference Based Measures. A critical overview of reviews. *Value in Health*. 2015. *18; 7: A744.*

Conference presentations

Finch AP, Brazier JE, Mukuria C. Testing the impact of candidate bolt-ons on preferences using pairwise choices. Paper presented in the plenary session at the 34th EuroQol plenary meeting, Barcelona, Spain, September 2017.

Finch AP, Brazier JE, Mukuria C. Selection of bolt-ons after factor analysis identification. Are linear regressions a useful technique? Paper presented in a parallel session at the 2nd Advances in Patient Reported Outcomes Conference, Oxford, United Kingdom, May 2017.

Finch AP, Brazier JE, Mukuria C. Selection of bolt-ons after factor analysis identification. Are linear regressions a useful technique? Poster presented at the 2nd EuroQol Academy Meeting, Noordwijk, Netherlands, March 2017.

Finch AP, Brazier JE, Mukuria C, Bjorner JB. An exploratory study on identifying bolt-ons using principal component analysis and structural equation modelling: The EQ-5D example. Paper presented in the plenary session at the 33rd EuroQol Plenary meeting, Berlin, Germany, September 2016.

Finch AP, Brazier JE, Mukuria C. An exploratory study on identifying bolt-ons using principal component analysis and structural equation modelling: The EQ-5D example. Paper presented in a parallel session at the Summer 2016 Health Economists' Study Group (HESG), Gran Canaria, Spain, June 2016.

Finch AP, Brazier JE, Mukuria C. Validity and responsiveness of preference based measures: A critical overview of reviews. Poster presented at the ISPOR 18th Annual European Conference, Milan, Italy, November 2015.

Abstract

In health economics benefits are often measured in terms of quality adjusted life years (QALYs), an index that combines the length of life and the health related quality of life (HRQoL) of patients. HRQoL is usually estimated using generic preference based measures (GPBMs). To ensure comparable decisions it would be recommended to use only one GPBM in all assessments, but this is not always appropriate due to validity and responsiveness problems of these measures. When this happens, bolt-ons represent a potential solution. Bolt-on research is at an early stage, and the process of how to identify and select bolt-ons is still unclear. To address this gap, this thesis examines methods for identifying and selecting bolt-on dimensions, using the EQ-5D as a case study.

This thesis summarizes the results of four studies. The first study examined the performance of the 5 most commonly used GPBMs across all disease areas and conditions, using an overview of systematic reviews. The second study investigated the possibility of using factor analysis techniques to identify bolt-on dimensions. The third study explored whether it is possible to select between bolt-ons using their ability to predict differences in HRQoL. The fourth study examined whether it is possible to select between bolt-ons using preferences elicited from pairwise choices over health states.

A number of important findings were made. First, GPBMs appear generally valid and responsive across many disease areas, but the quality, nature and breath of evidence makes difficult to draw definitive conclusions. Second, factor analysis techniques can be used to identify bolt-on dimensions. Exploratory and confirmatory factor analysis need to be used in conjunction. Third, bolt-ons differ in their ability to predict differences in HRQoL and for this reason they might be selected based on this information. Fourth, bolt-ons differ in terms of their impact on preferences and for this reason they might be selected based on this information.

Overall, a key contribution of this thesis is to systematically examine methods for identifying and selecting bolt-ons for generic preference based measures.

Chapter 1
Introduction to the thesis

1.1 Introduction

Health is a complex concept that is subject to change, related to the value system of societies and dependant on the historical period. For this reason, numerous definitions and conceptualizations of what health is have been proposed in the last century (e.g. Lewis, 1953; Parsons, 1958). Among them, the statement included in the constitution of the World Health Organization has been particularly influential. According to this statement, health is "A state of complete physical, mental and social well-being, and not merely the absence of disease and infirmity" (WHO, 1948).

Numerous measures of health have been developed in order to operationalize the WHO definition (Brazier et al, 2017). These measures differ in the dimensions and items they include, but have as a common denominator the recognition of the existence of a quality of life component to health that can be examined through different questions. For this reason, measures of health developed under the WHO framework have been commonly referred to as health related quality of life (HRQoL) measures, where HRQoL can be considered as "how well a person functions in their life and his or her perceived wellbeing in physical, mental and social domains of health" (Hays and Reeve, 2010).

Measurement of health is of interest for economics, as health care consumes a large proportion of every industrialized country resources (McGuire et al, 1988). Economics is concerned with resource allocation, assessing costs and benefits in order to identify which options maximize utility. The traditional approach to measure benefits has been using preferences translated into monetary values. However, for a number of ethical and empirical reasons financial metrics are often not used in the health care sector (e.g. Olsen and Smith, 2001), with quality adjusted life years (QALYs) being the preferred outcome.

QALYs combine survival and HRQoL of patients in a single index, calculating the latter on a scale from 0 to 1 that reflects the strength of preferences for that health state or condition (Drummond et al, 1987). A weight of 1 corresponds to optimal health, while a weight of 0 corresponds to a health state or condition judged to be equivalent to death (Gold et al, 1996). Three valuations methods are commonly used to elicit preferences, and these are the Visual Analogue Scale, the Time Trade Off and the Standard Gamble (Brazier et al, 2017).

Generic preference based measures (GPBMs) of health are a type of HRQoL measure that uses preference as a scoring system (Brazier et al, 2017). They comprise of a standardized descriptive system through which health is described, and a weighting system where values are elicited using preferences from members of the general public. There are several GPBMs which differ in the dimensions and items they cover and the methods they use to elicit preferences, generating substantially different values (Nord et al, 1993). For this reason, one would suggest that consistent assessments between programs and interventions should be ensured using the same measure. However, this is not always appropriate, as GPBMs might not be valid or responsive in some situations (Longworth et al, 2014). In those cases, possible alternatives are eliciting values directly from patients, using a different GPBM or using a condition specific measure (Brazier et al, 2012). All these solutions come at the cost of the comparability between assessments. Using a different generic or condition specific measure implies valuing HRQoL based on different dimensions of health and valuation methods, while using patients own valuations introduces sources of non-comparability such as response shifts (Post et al, 2001) and adaptation to the health states (Sprangers and Shwarz, 1999; Brazier et al, 2017).

Recently, a fourth option has emerged, which involves adding bolt-ons to the descriptive system of the GPBM deemed inappropriate in a condition or disease area due to missing key domains of health. This technique might represent a viable solution to strengthen measures' validity and responsiveness and to improve comparability between assessments. This is because the additional dimensions might help detecting relevant aspects of health for the condition or disease area of interest, while simultaneously maintaining the core descriptive system of the reference measure. However, little research has been conducted to date in the field, with the questions of how bolt-ons should be identified and selected remaining mostly unaddressed.

1.2 Aims and objectives

This thesis aims at examining methods that could be used to identify and select bolt-ons for GPBMs. Given its methodological focus, any GPBMs could have been used as a case study. The EQ-5D has been selected among the others because it is the preferred measure of the National Institute for Health and Care Excellence (NICE) in England (NICE, 2013) and because it has been widely used in economic evaluations worldwide (Szende et al, 2007). Moreover, the measure has the advantage of having a smaller descriptive system compared to other GPBMs, which eases the process of bolting on new dimensions.

This thesis will address a set of specific objectives to meet its aim:

1. To review the psychometric characteristics of the GPBMs commonly used in economic evaluations across disease areas and conditions and to review the bolt-on studies conducted to date;
2. To examine the usefulness of employing quantitative methods such as principal component analysis, exploratory factor analysis, confirmatory factor analysis and structural equation modelling for the purpose of bolt-on identification;
3. To examine whether it is possible to use the ability of bolt-ons to predict differences in HRQoL to select between them;
4. To examine whether it is possible to use the impact of bolt-ons on preference for pairs of health states to select between them.

1.3 Structure of the thesis

This thesis comprises of seven chapters. After the current introductory chapter, chapter 2 describes different conceptualizations of health and how the holistic model has been operationalized into measures of HRQoL. It introduces how resources are allocated according to mainstream economic theory and explains why government intervention is required in the health care sector. It presents economic evaluations as a means of informing decisions in the health care context and describes the differences between cost benefit analysis (CBA), cost effectiveness analysis (CEA) and cost utility analysis (CUA). It concludes providing an overview of the different approaches and methods used to elicit utilities for health states or conditions for CUA.

Chapter 3 focuses on GPBMs, the mostly used approach for obtaining HRQoL utilities commonly used in CUA. It describes the five mostly used GPBMs and explains that these differ in the dimensions they cover and the valuation methods they employ. It provides an overview of the existing literature on the psychometric performance of GPBMs across all disease areas and conditions. The overview summarises evidence from published and unpublished systematic reviews and extracts the information at a study level. The chapter also explores possible alternatives commonly used when the chosen measure is not appropriate in a condition or disease area with a focus on bolt-ons. A review of the bolt-on studies published to date is undertaken with a specific focus on methods used for identifying and selecting bolt-ons. It concludes that further examination of some of the methods used, and other possible alternative methods is required to understand whether they can be employed to identify and select bolt-ons. In doing this, the chapter addresses objective 1.

In response to the need to explore methods for the identification of bolt-ons for GPBMs, chapter 4 examines the potential of using factor analysis, therefore addressing objective 2. Firstly, the chapter introduces principal component analysis (PCA), exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and structural equation modeling (SEM) and explains why these techniques might be useful for identifying bolt-on dimensions. Methods for undertaking these techniques are described and their results provide a list of factors, and items related to them, which can be developed or adapted into bolt-on dimensions for the EQ-5D. Results are discussed taking into account the strengths and limitations of each approach.

Chapter 5 addresses objective 3 by investigating the possibility of selecting candidate bolt-ons identified through previous quantitative research using their ability to predict differences in a proxy of HRQoL. This chapter employs two tests. The first test examines whether bolt-ons are able to predict differences in HRQoL. The second test further examines whether these bolt-ons are able to account for differences in HRQoL between members of the general public and patients affected by nine chronic conditions. Results from the two tests are discussed including how they contribute to the selection of bolt-ons.

Chapter 6 addresses objective 4 by investigating the possibility of selecting candidate bolt-ons using their impact on preferences for pairs of health states. The chapter first describes the methods used for developing 8 bolt-ons from the 6 identified factors and how these were tested in terms of their face validity and relevance. It then sets out the methods for undertaking the pairwise choices. Results are presented and discussed.

Chapter 7 summarizes this thesis, presents its main findings, highlights their relevance and explains how they contribute to further the existing knowledge on bolt-ons. It describes the limitation of the thesis and it provides recommendations for further areas of research.

Chapter 2
Background

2.1 Introduction

The aim of this chapter is to provide the background for this thesis. It introduces 4 models of health and gives an overview of the concept of wellbeing. It explains how the operationalization of the holistic model, one of the four models of health described, resulted in a number of measure that differ in their content but all aim at measuring HRQoL. It presents the mainstream economic model of resource allocation and explains why the health care market cannot be left unregulated. It clarifies that health care decisions are informed using economic evaluations and presents three types of them, focusing on cost-utility analysis. It introduces QALYs, the outcome measure of CUA, a metric that encompasses the length of life and an index of HRQoL measured in terms of utility. Commonly used approaches to measure HRQoL are discussed, as well as the three preference elicitation techniques used to derive the utility associated to the health states used to describe the individuals' HRQoL.

2.2 Health and wellbeing

The meaning of health is complex, subject to change, related to the value system of societies, of interest for multiple disciplines and dependant on the historical period. It is therefore not surprising that numerous conceptualization and definitions of what health is (e.g. Lewis, 1953; Parsons, 1958; Mechanic, 1968; Wilson, 1970) and which aspects it includes (e.g. Kelman and Wilner, 1962; Ruesch and Brodsky, 1968; Nagi, 1969; Safilios-Rothschild, 1970; Jahoda, 1958; Allport, 1960; Smith, 1969; Clausen, 1968) have been proposed. Although there have been many different ways of conceptualising health, most of the last century conceptualizations and definitions relate to four main models of health: the medical model, the holistic model, the wellness model and the environmental model (Larson, 1991).

2.2.1 The medical model

The medical model is the reference conceptualization employed by the medical literature and defines health as the absence of diseases (Larson, 1998), where diseases are conditions of the body in which its structure or function is disturbed or deranged (Wood, 1986). As diseases are not directly observable, illnesses are used as a signal of the presence of a disease. Illnesses are perception of the individuals that they are suffering from a disease (Wood, 1986), and manifest themselves through symptoms. The medical model embraces a negative conceptualization of health, where ideal health is the absence of disease, and the purpose of healthcare is to eradicate these diseases (Mold, 1995).

The medical model has been criticised as the connection between diseases and illnesses does not account for the possibility of having asymptomatic diseases, or symptoms without having a disease (Larson, 1998). More importantly, the medical model ignores the fundamental role of preventive medicine and of the social determinants of a disease (Culyer, 1983).

2.2.2 The WHO definition (or holistic model)

The holistic model tends to equate with the World Health Organisation definition of health as “A state of complete physical, mental and social well-being, and not merely the absence of disease and infirmity” (WHO, 1948). This broader conceptualization includes aspects such as social participations and wellbeing as relevant and integral components of what health is, presenting a positive conception of health together with the more traditional negative one (Payne, 1983).

The holistic model has been criticised as its broadness has been considered a contributing factor to the medicalization of society (Garner, 1979). This point is well explained by Smith (2008, pg. 1), who argued that using the word complete “leaves most of us unhealthy most of the times". Critics have also expressed doubts on the holistic model considering it utopic and of difficult applicability in the real world (Larson, 1998). Some authors have argued against the inclusion of social wellbeing into the definition of health. For example, Ware and colleagues (1981) claimed that social factors are important aspects of health, but they should not be used to define the status of an individuals' health as they are not part of it.

2.2.3 The wellness model

The wellness model can be considered as a spinoff of the holistic model of the WHO definition of health. The model focuses on the wellbeing component of the WHO definition, and describes health as an "intuitive notion [made of] physical wellbeing, comfort, energy and ability to perform" (Greer, 1986). The model emphasizes the subjectivity of the individuals in self-judging their own health. This is well explained by Marvin and Crown (1976), who clarify that the wellness model describes health in terms of "a feeling or experience that people either enjoy or lack". The wellness model often includes notions of capability and potential achievement. For example Dunn, one of the most important promoters of this conceptualization, stated that wellness is "an integrated method of functioning which is oriented towards maximizing the potential of which the individual is capable" (1961, pg. 4-5).

A major criticism of the wellness model is that it expands the meaning of health in terms of happiness. This is seen as a limitation, as it is perfectly possible for an individual to be in full health, but still be unhappy (Bice, 1976). Another common criticism of the wellness model is that happiness is a very relative concept related to a number of factors, such as age and cultural contexts (Greer, 1986). Defining health in terms of happiness would result in having perfectly healthy individuals always considered as in need of care, as well as providing less care than needed to populations generally more happy.

2.2.4 The environmental model

The environmental model argues that a definition of health cannot exist if the interaction and relationship between the individual and its environment is not accounted for. Examples of health definitions connected to this model are those of Rosedale (cited by Navarro, 1977), who claims that health is “the product of an harmonized relationship between man and his ecology”, and Dubos (1968, pg. 69), who claims that health is “a modus vivendi enabling an imperfect man to achieve a rewarding and not too painful existence while coping with an imperfect world”.

A major criticism of the environmental model is that it ignores the individual, concentrating mainly on the environment (Larson, 1991). This opens to the possibility of having the same individual in the same state considered healthy in one environment and not healthy in a different environment (Larson, 1991).

2.2.5 What is wellbeing?

As noted, some definitions of health refer to wellbeing e.g. the WHO definition refers to wellbeing while the wellness model can include aspects of happiness, which falls under subjective wellbeing. It is therefore important to consider what is meant by wellbeing.

Wellbeing is a complex concept that can be broadly defined as how well an individual life is going (Brazier, 2017). In an effort to identify what makes a good life, scholars have adhered to four main schools of thought. These are: hedonic theories, eudemonic or flourishing theories, life satisfaction theories and objective list accounts.

Hedonic theories conceptualize wellbeing in terms of pleasure (Fava and Riuni, 2003). These theories argue that when pleasure and happiness increase, this is a signal that the individuals’ wellbeing is increasing. By contrast, when unpleasant and painful experiences happen, this shows a decrease in the individuals’ wellbeing. Eudemonic or flourishing theories conceptualize wellbeing in terms of someone’s ability to fulfil his or her own nature. In presence of higher degrees of autonomy, self-acceptance and positive relations, these theories argue that there are high degrees of wellbeing (Ryff, 1995). Sen's capability approach, that emphasizes the possibility of doing rather than the actual choice of doing, is an example of an eudemonic conceptualization of wellbeing. Life satisfaction theories claim that wellbeing is someone's assessment of his or her own life (Haybron, 2008). These theories argue that if someone declares that he or she is satisfied with his life, then his or her wellbeing is high. Objective list accounts attempt to identify objective characteristics that are important for wellbeing. These might include aspects such as literacy and accommodation (Brazier et al, 2017).

In the literature, distinctions have been made by referring to subjective wellbeing which is often taken to mean people's emotional responses or their satisfaction with domains such as health, income or relationships, and global judgments of life satisfaction or quality of life (Diener et al, 1999). These are related to some of the hedonistic and life satisfaction theories without having a distinct theoretical rooting other than in their being subjective.

Equally, there are links between wellbeing and quality of life. Moons and colleagues (2007) reviewed past attempts to define and measure quality of life, including objective indicators and subjective perceptions of quality of life. They noted that an agreement exist on the existence of both physical and subjective components of quality of life, with a growing perception that the subjective components might be better to define the term. Examples of this subjective orientation is the conceptualization of quality of life as happiness and satisfaction with life (Fayers and Machin, 2007; Meiselman, 2016).

There is a debate as to whether wellbeing is a distinct aspect from health or whether it is the key outcome. Some of the health measures that will be discussed more in detail in chapter 3 such as the Assessment of Quality of Life (AQoL 8D) explicitly attempt to incorporate more psychosocial concerns into their description of health (Brazier et al, 2017). Others, such as the Health Utility Index Mark 3 (HUI 3) cover aspects that are also emotional wellbeing items such as happiness. Equally, health may be seen as contributing to overall wellbeing (Diener et al, 1999; Ryan and Deci, 2001), where wellbeing provides information on the actual experiences of individuals as a result of the change in their health (Dolan and Kahneman, 2008).

These distinctions may be more salient in some conceptualisations of wellbeing than others.

2.3 Operationalization of the holistic model of health

Definitions and conceptualizations are useful in that they inform how to measure the concept of interest. This section focuses on the holistic model as this is closely related to preference based measure of health which are the focus of this thesis. Operationalizing multidimensional concepts such as health (in its holistic definition) is not easy. This is because multidimensional concepts cannot be measured directly. As health cannot be measured directly, indicators i.e. dimensions or components should be used to infer the overall index (Lerner, 1973). From the WHO definition it emerges that an index of health should include indicators tapping on three aspects: physical health, mental health and social health/ wellbeing. However, the definition does not clarify the conceptualization of these three aspects, nor which indicators are relevant or important (Lerner, 1973). In an attempt to better clarify health for measurement purposes, the WHO proposed a further explanation where health was described as a continuum where diseases cause impairments e.g. pain, which have an impact on abilities e.g. limited walking, that in turn reduce participation e.g. reduced usual activities (WHO, 1980).

Although helpful, the new classification did not resolve the issue of which elements of this continuum of health should be embraced for each of the three aspects of health i.e. physical, mental and social health (Fayers and Machine, 2007). More importantly, even agreeing on the same elements for each of the three aspects would not result in the inclusion of the same indicators. This is because different indicators i.e. dimensions or items might refer to the same conceptualization of the aspect of health of interest. For example, physical health in terms of functional abilities can be measured asking questions on an individuals' ability to move, but also asking questions on an individuals' ability to get around his house. These complexities have resulted in a large number of measures aiming at measuring health under the WHO framework, but using different dimensions and constructs. Authors and measures developers have made their own value judgments on which elements to include. For example, the RAND experiment conceptualized physical health in terms of functional abilities, mental health in terms of affectivity and self-control, and social health in terms of participation (Ware, 1980; Newhouse et al, 1993).

Notwithstanding these differences, a common denominator of these measures is the recognition of the existence of a quality of life component to health i.e. measuring health requires taking into consideration the length of life and the quality of life during that period, and of different health questions through which this component can be examined. Focusing on this similarity, scientific publications started to refer to measures of health developed under the WHO framework as health related quality of life (HRQoL) measures.

HRQoL is a loose term that until today continues to have an ambiguous and controversial meaning. In a review examining the use of the terms health, quality of life and HRQoL, Karimi and Brazier (2016) found that HRQoL was used in the literature with at least 4 different meanings, some of which resembled the concept of health, and some of which resembled the concept of quality of life. Although the issue of what HRQoL actually means is important, it is beyond the scope of the current introduction. However, what it is relevant to notice is that numerous measures of health have been developed following the WHO definition. They include a plurality of dimensions related to the three core aspects of health of the WHO but differ in the dimensions that they cover (Fayers and Machin, 2007). These are empirically referred to as HRQoL measures.

This thesis focuses on one type of HRQoL measures commonly used in the health technology assessment of healthcare intervention, namely generic preference based measures (GPBMs). Before these can be presented, it is important to describe the reasons why these measures are important in the context of healthcare decision-making. For this purpose, it is first necessary to provide a background to how decisions are made in economics, why markets cannot be left unregulated in the health care sector and what methods have been used to inform decisions given these premises. These issues are explained in the next sections of this introduction.

2.4 Allocation of resources in mainstream economics

Resource allocation is an important consideration in decision-making, as needs are unlimited but resources are limited. In neo-classical economics (or welfare economics) markets are thought to be the optimal method to allocate resources. Individuals are considered rational and fully informed agents able to compare alternative goods and services, to rank them consistently according to their preferences, and to choose between them in order to maximize their utility, of which they are considered to be the best judges (Hodgson, 2009). Information is free and easily accessible, and uncertainty is at a minimum. As a result of these elements, individuals can forecast the utility associated with different choices, and the utility they obtain by those choices coincides with the one forecasted. The individualistic conception of society of neo-classical economics determines a model where the overall societal wellbeing is a simple product of the individual utilities of the members of the society, the value judgement at the base being that trade-offs between individuals are treated as equal (Morris et al, 2007). As a consequence, the most efficient allocation of resources appears to be the one where any reallocation cannot make someone better off without making someone else worse off. This, known as the Pareto efficiency principle, provides a normative rule of which outcomes the society should prefer and the markets would reach if left unregulated (Busino, 1987).

2.5 Problems and solutions to neo-classical economics allocation of resources

Pareto efficiency model has been widely criticized. In a series of articles published in the Journal of Economic Perspectives, the issue of which and when Pareto improvements could be found in real world was investigated, with most authors concluding that rarely, if ever, did the model fit real decisions (Morris et al, 2007). Obviously, if welfare economics aims at providing a useful framework for understanding how to choose between alternative options this lack of applicability is problematic. In response to this issue Kaldor argued that, if after a change winners were able to compensate losers, everybody in the society would be better off or equally well as before and the change could be considered to be socially worthwhile. This development of the Paretian principle, known as the Kaldor-Hicks criterion

(Kaldor, 1939; Hicks, 1939; Hicks, 1941) has enjoyed widespread support by welfare economists (Drummond et al, 2007) and has been suggested also in the healthcare sector (see CBA below).

More relevantly, there are further criticisms of welfare economics which apply specifically to the healthcare sector. Healthcare consumers cannot know with certainty when they will need to access treatments and cures, because it is impossible for them to forecast if and when they will get ill. The fact that satisfaction in consuming health care is afforded only in the event of sickness determines an unsteady demand, which significantly departs from the demand seen for many other goods (Arrow, 1963). Moreover, consuming healthcare does not guarantee an improvement in health, which allegedly, is the ultimate reason why the good is consumed. In other words, the demand for health is derived from the demand of health care (Brazier et al, 2007). However, most people will need to access costly medical treatment at some point in their life, and consequently mechanisms to share the risk are required. This has led to the emerging of insurances, which became the prevalent method for paying for medical services (Pauly, 1986).

Healthcare insurance markets themselves present some problems. One that has received large theoretical and empirical attention (Pauly, 1986) is moral hazard. Moral hazard has been described as “the failure [of consumers] of holding up moral qualities” (Faulkner, 1960, pp.327) as a result of being insured, with a consequent increase in the consumption of good and services over the level it would be consumed if the cost had to be borne by consumers (Pauly, 1986). Moral hazard can take two forms: a reduction in the consumption of preventive care (ex ante moral hazard) or an increase in consumption when ill health occurs (ex post moral hazard) (Pauly, 1986). While the relevance of the first case, particularly emphasized in the principal-agent literature (Pauly, 1968), is debatable e.g. it is difficult to imagine people would normally reduce their attention when skiing as a mere effect of being covered for the cost of a possible injury, the second one represents a significant problem for insurance markets. Ex post moral hazard originates from an asymmetry of information, where insurance companies cannot objectively define the severity of the illness of the insured once the illness occurred (Pauly 1971; Newhouse 1978) and it has two main effects. Firstly, it results in a welfare loss (Nyman, 1999) that corresponds to the consumption of unit of medical care whose value to the consumer is less than their cost (Pauly, 1986). Welfare losses have been quantified in a number of empirical studies and appear to be always in the order of billions of dollars (e.g. Feldman and Dowd, 1991; Feldstein, 1973). Secondly, in presence of increased consumptions, premiums will necessarily be adjusted upwards, and this increase in prices might drive away some individuals with lower risks, generating deadweight losses (Culyer, 1991).

Another problem commonly discussed in the health insurance literature is adverse selection. Adverse selection generates when individuals that bear a greater than average risk of ill health do not fully disclose this information to the insurer (Pauly, 1986), accepting the average premium. Given that insuring these individuals increases the pooled risk, insurance companies will face losses and will be induced to adjust premiums upward. Lower risk individuals might be driven out of the market, with the consequence of having a large pool of uninsured individuals and a small pool of insured ones that sustain heavy financial burden (Culyer, 1991), or the impossibility of achieving an equilibrium (Pauly, 1974; Rotshild and Stiglitz, 1976).

The uncertainty surrounding the time at which medical treatment will be needed goes together with the uncertainty of the product consumed. Consumers of healthcare often cannot compare alternatives e.g. if they were victim of an accident or make rationale choices e.g. if they are mentally ill (Pauly, 1986), and even when their cognitive abilities are at the best, medical knowledge is extremely complicated. This makes it difficult for patients to appropriately forecast the actual outcome of the treatments and to make rational decisions if not assisted by a physician (Arrow, 1963). A common way in which the physician-patient relationship has been described in the economic literature is the one of the principal-agent, where the patient delegates authority to the physician, who uses it to act in the patients interests (e.g. Evans, 1984; Phelps, 1992). Through this contract, the market positively and independently resolves the asymmetry of information. However, health economists have long recognized that perfect agency relationship does not exist, and that this can only be used as a model to study deviations from the benchmark behaviour (Gafni et al, 1998). One of the mostly debated and controversial topics in this area (Reinhardt, 1985; Rice and Labelle, 1990; Feldman and Sloan, 1988) is whether, among these departures, there is a tendency of physicians to generate demand in response to fee changes, declining market shares (Labelle et al, 1994) or other individual interests. This phenomenon, called supplier induced demand, would arguably increase the demand for health care over the necessary level (Donaldson and Gerard, 1989; Rice and Labelle, 1990), and certainly over the level consumers would independently choose. In order to test for the phenomenon empirically, researchers have employed different methods, among which testing for the effect of physician supply on the volume of medical care utilization (e.g. Evans, 1974; Stano, 1985), the effect of physician supply on physician income (e.g. Redisch et al, 1981) the effect of physician supply on fee levels (Fuchs, 1978; Redisch et al, 1981), the effect of physician supply on initiation to treatment (Rossiter and Wilensky, 1983; Tussing and Wojtowycz, 1986) and others. These empirical tests generated conflicting evidence leaving a lack of consensus on the circumstances under which supplier induced demand might arise and its policy importance (Labelle et al, 1994), up to the point that some have argued that ideology more than evidence drives results (Reinhardt, 1985), with others questioning the feasibility of ever deriving unequivocal results (Fuchs, 1986; Pauly, 1986). Nevertheless, supplier induced demand remains one of the most discussed possible sources of market failures of the healthcare sector.

Healthcare markets are also characterized by externalities. These are positive or negative effects of an industrial or commercial activity that might affect other parties from those involved in the activity without being reflected in the market prices. Examples of these can be seen in vaccination programs, where someone's vaccination has a direct effect on the health of other members of the society (in terms of reduced risk of contracting that disease) (Donaldson and Gerrard, 1989). More complex forms of externalities might exist. For example, given the nature of the good in question, individuals might experience a benefit by only knowing that everyone in need gets access to the appropriate cures, although this does not directly affect their own health i.e. philanthropic behaviour (Culyer, 1991).

Market failures such as those described above violate the key tenants of the neo-classical framework, and more particularly the utility principle i.e. individuals rationally maximize their welfare by ordering options, individuals' sovereignty (individuals are the best judges of their utility), and consequentialism (Hurley, 2000). It has been argued that this requires government intervention (Buchanan and Tullock, 1962). A more theoretical criticism of welfare economics application to the healthcare market is its "welfarism", the idea that "the goodness of any situation should be judged only based on the utility levels attained by individuals in that situation" (Hurley, 1998, pg. 377) or, as put by Sen (1986, pg. 111) "judging the goodness of state of affairs only by utility information". Culyer (1991) reduced the scope of welfare economics by arguing that welfarist economics approach is highly restrictive, as considering social welfare only as a function of individuals' utilities, based on goods and services consumption, does not allow judgments on non-utility aspects of the world.

In response to this confined evaluative space, a different approach, named extra-welfarist, arose. Extra-welfarism transcends welfare economics by arguing that societal wellbeing should include aspects other than the simple aggregation of individuals' utilities, such as equity considerations, characteristics of individuals and capabilities (Brouwer et al, 2008). Building on this, the health economics tradition has argued that health should be considered as an independent argument in the welfare function, opening to a new notion of efficiency, where value judgments are explicit and are incorporated into the maximand, which is an index of health (e.g. Culyer, 1991). Simultaneously, the possibility of including equity considerations is envisaged, with a process of mediation between equity and efficiency being required in order to guarantee that societies can reach an endowment point from which an acceptable distribution of resources is possible (Folland et al, 2007).

2.5.1 Summary

This section has described the neoclassical theory of allocation of resources and the problems in leaving the market unregulated for the healthcare sector. In doing this, it has also introduced the reader to the welfarist and the extra-welfarist economic traditions. Once it has been determined that allocation decisions should not be left to the unregulated market forces, a need for decision criteria arise (Brazier, 2007). Economic evaluations represent the main tool through which decisions have been informed for the healthcare sector. These techniques build on the two economic traditions previously described i.e. welfarist and extra-welfarist, and choosing one underlining philosophy over the other has repercussions in the way outcomes are measured (Brazier, 2007). This will be discussed in the next section of this thesis.

2.6 Economic evaluations in healthcare decision-making

In the real world societies have scarce resources. The proliferation of new drugs and medical devices pressurizes the limited budgets (Walker et al, 2011), leaving governments with the difficult decision of which services to provide, to whom, how, where and when (Drummond et al, 2005). Choices have important consequences for the individuals in the society (Brazier et al, 2007). Allocating resources to a vaccine program for all children under five in a given location might imply that other programs for the same population cannot be implemented, or that the same program cannot be delivered to different populations i.e. older populations, or that different programmes cannot be provided to different populations. In other words, in every health care decision the inherent lost opportunity of not providing an alternative option needs to be considered (economists commonly refer to this as to the opportunity cost of the decision) (Brazier et al, 2007).

In taking decisions, governments operate pursuing their policy goals. Some health economists claim that efficiency is the main one of them, as the desire of deploying resources efficiently to maximize community's health is the natural outcome of any health system (Wagstaff, 1991). This point is clearly explained by Drummond (1989, p. 71) who suggests that governments are concerned "with economic efficiency rather than notions of equity or social justice". Nevertheless, the issue is not undisputed, with some other authors reasoning that "notions of fairness, opportunity, freedom and rights are arguably of more importance in policy making than are concerns of moving individuals up their given preference ratings" (Hausman and McPherson, 1993, pg. 676). Certainly, many health systems should be explicit in explaining how resources are allocated (Brouwer et al, 2008), and often equity considerations are taken into account. For example, in the United Kingdom the 1944 white paper on health "A National Health Service" states that "the government […] wants to ensure that in the future every man, woman and child can rely on getting […] the best medical and other facilities available; that their getting them shall not depend on whether they can pay for them, or any other factor irrelevant to real need" (as quoted in van Doorslaer et al, 1993). Similarly, the Sozialgesetzbuch, the German social code, states that "medical care should be provided solely according to an individual's need, whereas the financing of care should be based solely on the individual's ability to pay (Pfaff and Wassener, 2000, p. 907). For the purpose of this introduction, the debate on the relative importance of equity and efficiency for priority setting is of little relevance. However, it is worth noticing that these are both important policy goals for many health systems, and that sometimes trade-offs between them are needed (James et al, 2005).

Various techniques have been proposed for promoting the aims of an efficient and equitable use of health care e.g. case based payment systems; budget caps per disease areas etc. Among them, economic evaluations have a crucial role in that they are able to describe, at least partially, the benefits of a given intervention compared to the opportunity cost of choosing the next best option, thereby addressing efficiency and equity aims. Broadly, economic evaluations are a series of analytical approaches used to clarify the possible alternative courses of action through the comparison of both costs and consequences of different interventions and programmes (Drummond et al, 2005). Although economic evaluations methods can be highly technical and involve the use of statistical techniques and modelling, at their core there are views of what are benefits and how these should be valued (Morris et al, 2007). In addition, different economic evaluations are grounded on value judgements coming from different economic traditions. Three main types of economic evaluations have been used in health care decision-making (Drummond et al, 2005): cost-benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA). If on the one side the measurement of costs is similar across these (e.g. Drummond et al, 2005), on the other side substantial differences exist in the measurement and valuation of the effects.

2.6.1 Cost-benefit analysis

CBA compares the costs of a program or intervention with its benefits, where these are valued in financial terms (Kalarman, 1974; James and Stokes, 2006). Because both gains and losses are calculated using the same metric (monetary), CBA can be considered as the direct application of the Kaldor criterion. In fact, as in the Kaldor criterion a socially advantageous improvement has some winners, and losers that can be compensated by part of the gains of the winners, also in CBA benefits and losses are directly compared in the form of a net benefit (Buchanan and Wordsworth, 2015). If the net benefit is positive (benefits are larger than costs) the program represents a better allocation of resources compared to the status quo (Drummond et al, 2005). In addition, given that benefits are calculated in monetary terms, CBA permits to compare returns in investments in different activities (Robinson, 1993), for example health and education. CBA employs a perspective where the patient is the consumer of healthcare, and uses its values to assess the impact of interventions (Brazier et al, 2017).

One of the first approaches used for valuing benefits in CBA has been the human capital. In the human capital individuals are considered, similarly to capital equipment, able to generate a flow of productive activities, the value of which corresponds to the rate of pay for every activity (Robinson, 1993). Morbidity or mortality costs can be estimated summing the value for the activities the individual would have performed in absence of a disease, which is the sum of lost income per year. The use of the human capital approach in health care has raised three main criticisms. First, the idea of putting a value on human lives is controversial and considered by many as strongly unethical (Aldred, 2009). Second, using rates of pay as a measure of benefit neglects those benefits that accrue to individuals not in the labour market such as reduction in suffering and pain (Robinson, 1993). Third, and from a purely economic perspective more importantly, the human capital approach is inconsistent "with the basic rational of economic calculus used in cost-benefit analysis" (Mishan, 1971, p.691) that requires the use of measures based on individuals' preferences. These criticisms have led to progressively abandon this method in favour of willingness to pay (WTP) ones.

WTP, as argued by Gafni (1991), are theoretically stronger as they estimate what individuals would be willing to pay to secure the implementation of a program, which is a direct extension of CBA theory. WTP measurements can be inferred indirectly from decisions individuals made (revealed preferences) or directly asking responders their preferences (stated preferences) (Johannesson, 1996). Indirect methods examine situations where a money/health outcomes trade off has occurred (O'Brien and Viramontes, 1994). For example, studies as these might infer the value associated to forms of insurance that reduce the risk of ill health in high risk individuals by analysing real world decisions of workers with a known health risk in accepting a raise in health premiums to cover that risk (Marin and Psacharopoulis, 1982). By contrast, direct measurement methods use surveys, questioning values for some phenomenon produced (or displaced) by the treatment, program or intervention under examination (O'Brien and Viramontes, 1994). For example, individuals might be asked, through an hypothetical scenario, the amount they would be happy to pay to reduce the risk of a medical treatment (Gafni, 1991). Direct WTP studies have been less employed than indirect methods, as real world situations where these are applicable are rare (Robinson, 1993). Indirect methods, also known as contingent valuation studies, use a variety of questionnaire types, which can be broadly categorised into open ended and closed ended studies (Mitchell and Carson, 1990; Johannesson and Jonsson, 1991).

Despite being theoretically stronger than the human capital approach, WTP methods might suffer from a number of problems. Effect biases, which arise when responders change their stated WTP amount for strategic reasons or in an attempt to comply with the expectations of the interviewer, might be seen in open-ended WTP studies (e.g. Carson, 1991). More relevantly, open-ended WTP studies appear to determine large numbers of non-responders due to the significant cognitive effort they require (O'Brien and Gafni, 1996). By contrast, closed ended studies are affected by starting point bias, where the maximum WTP of responders is influenced by the first bid presented, and range bias, where the range of presented amounts influences the maximum WTP (e.g. Johannesson et al, 1996; O'Brien and Gafni, 1996; Klose, 1999). Biases due to differences in designs are of crucial importance, as they can threaten the comparability of WTP estimates derived from different studies (Klose, 1999). However, two other biases make WTP particularly unattractive for the healthcare sector (Cookson, 2003). The first one is the insensitivity of WTP to the magnitude of the benefits (Kahneman and Knetsch, 1992; Baron et al, 1996; Bateman et al, 1997). An example of this case can be seen in a study by Beattie et al (1997) who found, using a high quality contingent valuation survey design, a tendency of responders in stating roughly the same monetary amount for any given reduction in the risk of death or injury. This is a major problem, as exaggerating the monetary value of reductions in small risks has the effect of inflating the importance of health care interventions that yield small benefits (Cookson, 2003). The second one is the tendency of WTP studies to inflate the value of the intervention under examination over the interventions not examined (Hausman and Mc Pherson, 1993). In other words, presenting responders with interventions in isolation increases their

WTP for that intervention, compared to when the same intervention is presented in relation to others (Cookson, 2003). This, also known as "budget constraint bias" (Mitchell and Carson, 1990), might lead to summing up WTP amounts well above the available budget.

These issues have led health economists to openly discourage the use of CBA in health care, as it is clear in Cookson (2003, pg. 892-893) who stated: "By generating inflated […] valuations, use of WTP methods by bodies such as NICE could play into the hands of the pressure groups and hamper efforts at rational decision-making. […]. In the context of resource allocation decisions within the health care sector […] economists and health care payers alike would be well advised to treat claims about the advantages of willingness to pay methods and CBA with scepticism". Similarly, Smith and Sach (2009, pg. 863) affirmed that despite having no restriction on the range of benefits valued and addressing allocative efficiency are potentially selling points of CBA, "the theoretical benefits of contingent valuation continue not to be realised, and sadly, Olsen and Smith's conclusion appears as relevant today as it did in 2001: contingent valuation remains on a road to nowhere in health economics".

The direct connection between CBA and the welfarist tradition has led some economists to support this technique in health care (e.g. Gafni, 2006; Johnson, 2012). Nevertheless, CBA remains less employed than other economic evaluations such as CEA and CUA, with most academics preferring to refine CEA methods rather than developing WTP ones (Gold et al, 1996; Drummond and McGuire, 2001).

2.6.2 Cost-effectiveness analysis

CEA compares the cost of a healthcare program to the natural unit of that program (James and Stokes, 2006). Possible examples of these natural units are the life years saved in a cancer treatment, the number of successfully treated patients or a patient reported outcome measure score i.e. not based on utility. A direct comparison of benefits and costs is not possible, but the analysis is well suited for comparing alternative treatments that have the same unidimensional goal (Drummond et al, 1987). In addition, it avoids CBA problem of measuring health in monetary terms (Johannesson and Jonsson, 1991).

CEA has received the favour of extra-welfarists because, differently from CBA, it focuses entirely on the maximization of health (benefits are measured as changes in health), excluding other components of utility from the analysis. However, the technique does not allow comparing treatment/programs/interventions across disease areas and disease groups (being focused on the natural unit of measurement only interventions with the same outcome can be compared) (Drummond et al, 2005), nor to compare interventions that differ in more than one treatment outcome (e.g. Brazier et al, 2007). For these reasons, CEA can be used only if the objective is to increase the level of effectiveness while maintaining a given budget, or maintaining the same level of effectiveness while reducing the budget (Morris et al, 2007).

2.6.3 Cost-utility analysis

CUA is a form of CEA that instead of comparing costs with the natural unit of the program compares them using a synthetic index of the length and HRQoL of individuals, with the later being based on utility. This is usually expressed in terms of QALYs and is easily obtained by multiplying the length of time passed in the condition by a value calculated on a scale from 0 to 1 (Drummond et al, 2005). This value is meant to reflect the strength of preference of individuals over different health states (utility). CUA shares with CEA the idea that there is no additional objective to health care than health maximization (Mooney, 2003) and in that it is favourably looked at by extra-welfarists. In addition, this seems the most reasonable approach in a number of health systems, such as the UK one, where the majority of individuals support public institutions expecting them to maximize health under the national budget (Garrison Jr., 2009). In fact, CUA employs a perspective where social welfare is externally assessed by the community, and uses values of an intervention derived from a representative sample of the general population (Brazier et al, 2017). Measurement and valuation of health for incorporation in QALYs are discussed in the next section.

CUA has the distinctive advantage of being able to provide comparative evidence between different disease areas and disease groups, thanks to the use of a broader measure of benefit. In addition, CUA has also the advantages of being able to compare interventions with more than one health outcome (Brazier et al, 2007), which is something usually not possible with CEA. Finally, using a threshold (see later point), CUA allows to make allocative efficiency decisions, which are not possible using CEA. These characteristics make this type of analysis the preferred economic evaluation for a number of health technology appraisals bodies. NICE, for example, explicitly states in the Guide to the methods of technology appraisal 2013 that: “For the reference case, cost-effectiveness (specifically cost–utility) analysis is the preferred form of economic evaluation” (NICE, 2013, pg. 37). Similarly, the Haute Autoritè de Santè in France, affirms in the guide for choices of methods for economic evaluation that “cost benefit analysis is not recommended in the reference case analysis […]. The two methods recommended by HAS are cost-utility analysis and cost-effectiveness analysis” (HAS, 2012, pg. 17-18). Also the General guidelines for economic evaluations from the Pharmaceutical Benefits Board of TLV in Sweden reports that: “Cost-effectiveness analysis is recommended, with QALYs as the measure of effect” (TLV, 2003, pg. 1). Despite the widespread use of this method, a problem of CUA is that, in absence of a direct comparability between health outcomes and costs, a threshold of the cost per QALYs that represents an improvement in social welfare is needed. Different methods have been proposed to set this threshold, with the prevalent approach at the moment being setting the threshold to exhaust an exogenously determined budget according to the production possibilities of the health care system (Culyer, 2002).

2.7 Quality adjusted life years

At the heart of CUA there are the QALYs. QALYs are a measure of health outcomes where a utility weight is assigned to each period of time in which the individual is alive. The utility can be defined as the desirability or preference for a given health state (combination of problems) or condition in which an individual is found (Torrance, 1987). This weight lies on a scale from 0 to 1 and corresponds to the HRQoL during the investigated period. A weight of 1 represents optimal health and a weight of 0 corresponds to health states judged to be equivalent to death (Gold et al, 1996), although health states considered worse than death exist and these receive negative weights.

2.8 Approaches to derive utilities for economic evaluations

Different approaches can be used to derive utilities for use in CUA. One of them consists in eliciting utilities directly from patients. Using patients' own valuation avoids the problem of having to describe a health state or condition, as the patient is experiencing it directly (Brazier et al, 2017). Despite this advantage, direct assessments of patients own health state is rarely employed as consensus has formed around eliciting preferences from members of the general public.

This decision is based on a number of ethical and practical considerations. Among the ethical considerations, it has been claimed that utilities should not be derived from individuals with vested interests and that ex ante preferences should be used to value health states or conditions as these reflect a view of public funding as a form of insurance (Brazier et al, 2017). Among the practical considerations, it has been seen that patients' generally report higher values for the same health states or condition (Murphy et al, 2001) as a result of adaptation (Sprangers and Shwarz, 1999; Brazier et al, 2017), and that patients values are often affected by response shifts due to the change in their health status (Post et al, 2001).

Once a decision has been made that preferences should be elicited from members of the general public, some form of description of the health states or conditions is required. One approach that has been used for this purpose consists in using vignettes. Vignettes are descriptions of a health state or condition. They are generally in the form of a narrative text or a bullet point list, although more sophisticated formats have been proposed, such as providing spectacles that reproduce visual problems (Abellea and Tsuchiya, 2007) or presenting videos with actors describing a health state (Lenert et al, 2004). Vignettes have been used in numerous conditions, among which asthma, depression and cancer (Lloyd et al, 2008; Revicki and Wood, 1998; Tolley et al, 2013).

Vignettes are a useful technique, but they are affected by some important limitations. First, concerns exist on the fact that vignettes generally describe conditions in terms of the most typical problems encountered by patients, and for this reason they are not able to reflect the varied and full distribution of outcomes in a population or patient group (Peasgood et al, 2010). Second, there are no guidelines on how to conduct vignettes studies, which results in vignettes having completely different formats. Third, the comparability of vignette-based utility is limited as their content is often tied to a specific outcome. Finally, as vignettes describe a particular health state or condition, they have to be developed ad hoc every time a new condition or health state needs to be investigated, a process that is expensive and time consuming.

A solution to these problems is provided by the use of standardized descriptive systems. Standardized descriptive systems are able to describe multiple health states using the same set of dimensions, items and levels that allow. They can be generic, describing outcomes relevant for multiple different conditions, or condition specific, in which case they cover all outcomes relevant for a specific disease. When an algorithm is attached to a standardized descriptive system to obtain utility values for each of its health states, the measure is defined generic or condition specific preference based measures.

While condition specific measures represent a valuable alternative to GPBMs, utilities are generally obtained using this latter measure. This is because of a number of advantages of GPBMs over condition specific preference based measures: i) they allow to omit the name of the investigated condition, which has been seen problematic as it might lowers the utility values obtained; ii) they avoid focusing effects due to a condition being presented in isolation; iii) they are able to cover a broader spectrum of health problems and therefore they can be sensitive to side effects and co-morbidities (Brazier et al, 2017). This thesis investigates GPBMs. Before introducing them in detail, however, it is important to describe the valuation techniques that are commonly used to elicit utility values which form the basis of GPBMs utility algorithm. These are presented below.

2.9 Valuation techniques

Three methods have been commonly used to derive utilities for economic evaluations, whether using patients' own valuations, vignettes or standardized descriptive systems. These are the VAS, the TTO and the SG. In addition to these methods, ordinal response data have been employed to estimate cardinal values for descriptive systems, with discrete choice experiments (DCEs) being the most frequently employed.

2.9.1 Visual Analogue Scale (VAS)

VAS is the easiest way for valuing health states. The scale consists of a line on a page with clearly defined values, where the best score is usually 100 and the worst score is usually 0. Scale intervals are important as VAS is conceptualized as a cardinal measure of preferences i.e. if the difference in desirability between outcome A and outcome B is twice as great as the difference in desirability of outcome C and D, the interval between A and B should be twice as large as the interval between C and D (Drummond et al, 2005). Responders are asked to value their own health on the scale, or health states described through a vignette or a standardized descriptive system. When multiple health states need to be valued, responders might be asked to locate all of them on the thermometer, choosing a distance between them that reflects their relative preference for the states.

VAS presents numerous variants. The length of the line might vary between versions, the scale can have different numbers, different middle points or different graduation points, and the thermometer layout can be vertical or horizontal (Paul-Dauphin et al, 1999). Descriptors (of both health states and anchors) need to be clear and unambiguous, as this guarantees comparability between responders (Brazier et al, 2007). However, anchors such as "best imaginable health" / "worst imaginable health" (Parkin and Devlin, 2006) are commonly seen in the literature, which might represent a problem as different responders might understand them differently in light of their own health status. This phenomenon has been named "scale recalibration" (Arons et al, 2013). For use in economic evaluation, anchors should be linked to the objective of measuring utility on a scale between dead and full health. It might happen that states worse than death need to be valued, in which case the standard descriptors "best imaginable health" and "worst imaginable health" could be used, but responders would be required to value death along with their own health, or the hypothetical health states presented in the vignettes (Brazier et al, 2007). This technique has been applied for example in the evaluation of the Quality of Well-being multi-attribute utility scale (Brazier et al, 2007).

VAS is an extremely practical instrument (Killewo et al, 2010) as demonstrated by its high completion and response rates (e.g. Torrance, 1976; Torrance, 1987; Gudex, 1996; Bakker et al, 1994). In addition, it appears substantially cheaper than other valuation techniques and quicker to complete (e.g. Torrance, 1976; Wolfson, 1982; Green et al, 2000). However, its use for deriving utilities has been strongly criticised by some health economists. The instrument, differently from other elicitation techniques, is based on psychological theories of response to sensory stimuli and not on economic or decision theory (Parkin and Devlin, 2006). The absence of a notion of choice, and therefore of opportunity cost or risk, has led some authors to cast doubts on its ability to reflect the strength of preferences on a cardinal scale (Johannesson, 1996; Killewo et al, 2010; Brazier et al, 1999). In empirical studies, some authors have noticed the presence of response spreading (Parducci, 1974; Kuppermann, 1997; Kaplan, 1993), a phenomenon where responders put states that are similar at some distance on the thermometer. Other authors found context effects, which consist in individuals' giving a value to a health states which is conditioned by the other states presented in the exercise (Bleichrodt and Johannesson, 1997; Nord, 1991; Sutherland et al, 1983). Finally, "end state aversion" biases, the tendency of responders of not using the end of the VAS scale, have been noted (Torrance et al, 2001).

2.9.2 Time trade-off (TTO)

TTO was developed specifically for use in health care by Torrance et al (1992) for states better than dead. It presents responders with a choice between two options, health state A for a time t and then death or a state B in full health for a shorter time t2 than t, followed by death. The time spent in full health t2 is varied until the responder is indifferent between the two alternatives, at which point the utility value for the intermediate health state is given by the ratio of t2/t. For states worse than dead, the TTO can be modified asking responders to choose between alternative 1, immediate death, and alternative 2, spending a length of time y in the state worse than dead followed by x time in full health, where x + y = t. Time y is varied until the responder is indifferent between the two alternatives, at which point the value for the state worse than dead is given by –x / (t-x) (Brazier et al, 2017).

TTO has numerous variants. Although most studies use 10 or 20 years timeframes (Arnesen, 2005) other durations have been employed, such as actuarial life expectancies (e.g. Stiggelbout et al, 1995; Essink-Bot et al, 2007) or responders own life expectancies (e.g. Heintz et al, 2013; van Nooten et al, 2009). Also, the mode of administration might vary between versions. The preferred method is face-to-face interviews, as the interaction between interviewer and responders generates data of the highest quality (Attema et al, 2013). However, face-to-face interviews are expensive and for this reason it might not be possible to use them in large studies. Hence, the use of alternatives such as online or postal surveys have emerged (e.g. Versteegh et al, 2013; Bansback et al, 2012b). These techniques are substantially cheaper, although interviewers cannot monitor the degree of effort put by the responders in answering to the task (Attema et al, 2013). Differences in the iteration procedure have been found in the literature. Some studies have used a bisection approach, which consists in dividing the difference in time in half till an indifference point is reached (Attema et al, 2013). Others, a top down or bottom up titration, where 1 year incremental or detrimental reductions are used until an indifference point is reached (Delquie, 1997). A third group of studies used ping-pong techniques, which consists in presenting durations at the top and bottom of the spectrum in consequent order (Lenert et al, 1998). TTO might also differ in terms of the visual aids used for assisting the task e.g. boards, cards or other graphical illustration (Attema et al, 2013) and the warm up tasks used e.g. questions with different health states, valuations of the same health state using different methods etc. (e.g. Versteegh et al, 2013; Dolan et al, 1996; Furlong et al, 1993).

TTO is an acceptable and practical method of health state valuation, as it has been demonstrated in a number of empirical studies (e.g. Ashby et al, 1994; Detsky et al, 1986; Patrick et al, 1994). The approach includes a notion of choice and opportunity cost, which led some to argue that it is grounded in consumer theory (e.g.

Johannesson et al, 1994; Parkin and Devlin, 2006). This claim has been questioned by a number of other authors. According to Mehrez and Gafni (1994), TTO cannot be considered as derived from consumer theory as it lacks a notion of uncertainty, which also makes it inappropriate for medical decision making. Dolan and Gudex (1995) and Green et al (2000) argued that the constant proportional time preferences assumption contradicts consumer theory, as individuals might have different time preferences e.g. individuals might prefer to incur ill health in different moments in their lives. For example, in presence of positive rates of time preferences, they might give greater weight to life years in the near future. Constant proportional time preferences have also been criticised because individuals' preferences might change while experiencing a condition (e.g. Sackett and Torrance, 1981). Individuals might, for example, adapt to a health state and therefore value it as less severe after time passed in it, or might become increasingly intolerant towards their condition (Brazier et al, 1999). Although issues on duration might affect all techniques of health state valuation, this problem is particularly relevant in TTO, as values cannot be thought to hold for states lasting for different time period (Brazier et al, 1999). In other words, the values given to different states cannot be considered independently from the duration of time passed in those states (Brazier et al, 1999), an important "framing effect" of this technique (Nord, 1992). A final criticism raised towards TTO is the fact that some responders might not want, for cultural or ethical reasons, trade any time of their lives to improve their quality (Scott, 1998).

2.9.3 Standard Gamble (SG)

SG presents responders with a given intermediate health state (for states better than dead), and a treatment with probability of success p (full health) and probability of death 1-p. The responders are given the possibility of living in the intermediate state for x years and then die, or to take the treatment. The probabilities of success or failure of the treatment are varied until the responder becomes indifferent from taking the treatment or being in the intermediate health state. The indifference point probability (p) is the utility associated to the intermediate state (Brazier et al, 2007). For states worse than dead the SG can be modified asking responders to choose between alternative 1, a treatment with probability p of return to normal health and additional t years of life in full health, or probability 1-p of remaining in the state worse than dead for t years and then die. Alternative 2 has the certain outcome of death. The probability p is varied until the responder is indifferent between the treatment and certain death, at which point the value given to the state worse than dead is given by –p/(1-p) (Brazier et al, 2017).

The SG has some variants, the main differences being the mode of administration and the procedure used to identify the indifference point. As regards the mode of administration, like TTO, SG can be self-administered online or by postal survey or interviewer-administered face-to-face or by telephone surveys (e.g. Glik et al, 2005). As regards the method to reach the indifference value, a widely used variant has been developed by Torrance (1986). This uses a probability wheel to help responders identify their indifference point. The probability wheel is an adjustable disc visual aid composed of two sectors, one for each possible outcome of the gamble alternatives. The probability of the two outcomes is proportional to the similarly coloured areas and is adjusted during the exercise (Brazier et al, 2007). An alternative variant of the SG, that uses titration, has been designed by Jones-Lee et al (1993). Responders are presented with a list of values for chances of success or failure of the gamble in a top-down or bottom-up format. Responders are asked to identify all the values where they would choose the treatment and all the values where they would reject the treatment. They are finally asked to choose the value where they would find more difficult choosing between the treatment and remaining in the described health state (Brazier et al, 2007). Ping-pong titration has also been used with SG.

Many economists consider the SG the preferred method for deriving utilities, given that it is directly drawn from the von Neumann and Morgenstern axioms of expected utility. Due to the uncertain nature of medical decision-making the method has been often portrayed as the "gold standard" for utility estimation (Drummond et al, 2005). This claim has been criticised by a number of scholars, with the main criticism being that the axioms of expected utility theory are routinely violated in real world decisions (e.g. Buckingham et al, 1996; Richardson, 1994). Also empirical concerns in the use of SG have been expressed. Some authors have found the method to be cognitively complex for some responders (e.g. Frobergand and Kane, 1989; Revicki, 1992; Stiggelbout et al, 1994), although no general pattern in this sense emerged from the literature (Brazier et al, 1999). More relevantly, some studies evidenced a tendency in responders of SG exercises to construct preferences while performing the task (e.g. Slovic, 1995). This "framing effect" is a major problem, as it casts doubts on the actual validity of values elicited through the valuation technique e.g. preferences might vary depending on the way in which questions are posed (Slovic,1995). However, as previously explained, framing effects are present also for other elicitation techniques e.g. TTO (Attema et al, 2013) and are not a specific negative feature of the SG. Another frequent criticism is connected to the influence that individuals' risk attitudes have on the values elicited (Kahneman and Tversky, 1979), as different individuals have different propensities towards risks (Bakker et al, 1994; Wakker and Stiggelbout, 1995) and these might alter the values elicited. For example, extreme risk aversion might lead some responders to be unwilling to accept any level of risk to obtain an improvement in health (Brazier et al, 1999), which might limit SG employment in some subgroups. By contrast, risk-loving behaviours might lead individuals to elicit health state values through SG lower than they would with different techniques as an effect of being able or enjoying to bear additional risks. A final source of complexity derives from the fact that these risk attitudes tend not to be constant in time but rather depend on the moment or the task performed (Loomes and McKenzie, 1989).

2.9.4 Discrete choice experiments (DCEs)

Discrete choice experiments (DCEs) present responders with pairs of alternative health states described in terms of attributes and levels of severity and ask them to express their preference between them generating in this way ordinal data. A subsequent phase, based on conditional logit models, is used to derive health state values. The use of DCEs for deriving health state values has been limited in the past as DCE preferences and coefficients are measured on a latent unobservable scale, which poses the problem of them not being anchored to the full health-dead scale needed for economic evaluations (Mulhern et al, 2012). Some methods existed to adjust for this (Rowen et al, 2011), but these were complex and time consuming, therefore preventing DCEs diffusion. More recently, Bansback and colleagues (2012) validated a novel approach that, using duration as an attribute, allows the user to obtain utility values directly. This approach has been rapidly adopted (e.g. Mulhern et al, 2014; Viney et al, 2014) and has been named DCE_{TTO}.

DCEs data can be collected using face to face interviews, postal or online surveys, or using computer assisted personal interviews. This latter method involves conducting face to face interviews but recording the answers using a computer tool. Each of these modes of administration have their strengths and weaknesses. Face to face interviews report good completion rates and generate high quality data, but are expensive and time consuming (Dolan et al, 1996). Postal surveys are cheap and easy to conduct, but response rates are often low. Online surveys offer flexible design opportunites and sampling options, but are prone to poor quality data and self-selection biases (Mulhern et al, 2013). Computer assisted personal interviews produce high quality data and offer flexible design options, but are more expensive than online administration (Mulhern et al, 2013). Differences between DCEs exist also in terms of the type of randomization performed e.g. randomization of the side, randomization of the order of the attributes etc., the methods used to select the health states valued e.g. full factorial design, D-optimal design etc. and the task presentation e.g. direct comparison of only two alternatives, comparison of multiple alternative simultaneously etc. (Mulhern et al, 2016).

DCEs include a notion of choice and are grounded in consumer and utility theory (e.g. Louviere et al, 2000). Among the empirical advantages of this method there are the ease of administration and the ease of comprehension for responders. This latter aspect has been argued to reduce the cognitive burden of eliciting health state values compared to other valuation methods (Brazier et al, 2017), a relevant strength particularly in low literacy and numeracy settings (Brazier et al, 2017). Moreover, standard DCEs without the use of a duration attribute have the advantage of not being affected by risk aversion or time preferences (Brazier et al, 2017). However, similarly to other techniques such as SG and TTO, DCEs might be susceptible to framing effects e.g. small differences in questions wording, context and mode of administration might result in large differences in responses (Brazier et al, 2017). A specific disadvantage of DCE using a duration attribute is the need of collecting data on a large number of combinations of attributes and health states to increase the certainty in the estimates obtained (Bansback et al, 2012). Moreover, similarly to

TTO, DCE_{TTO} are also affected by the fact that different individuals might have different time preferences.

2.9.5 Summary

In summary, different methods have been proposed to elicit values to use in economic evaluations. All these methods present a number of variants and differences that might alter the values produced. There is no consensus on which among them should be preferred, and although some have argued in favour of the SG from a theoretical viewpoint (e.g. Torrance and Feeny, 1980) others have criticised this claim (e.g. Richardson, 1994). In order to determine whether one of the methods should be preferred based on empirical evidence, Bleichrodt and Johannesson (1997) investigated these elicitation techniques relative performance comparing their results to preferences derived by direct ranking. The authors found that the TTO tends to respond more closely to direct ranking than the other two methods and concluded that TTO should be the preferred option. However, all these methods have been used for eliciting patients health, vignettes values and standardized descriptive systems value sets. For the purpose of eliciting health states of descriptive systems a method that has been proposed is DCEs. This latter has been recently used for the estimation of the EQ-5D-5L value function (e.g. Stolk et al, 2010).

2.10 An alternative to HRQoL: Subjective wellbeing

In recent years, there has been a rising interest in using subjective well-being (SWB) for economic evaluations. A number of articles investigating the relationship between HRQoL and SWB have been published, endorsing an increased role of SWB in public policy (e.g. Dolan, 2008 a; Dolan, 2008 b). These ideas have spread rapidly and have been accepted by some governmental agencies, as it can be seen from the all-party parliamentary group report on well-being economics that declared: "Subjective well-being evidence should be used in the calculation of QALYs to better inform the allocation of scares resources for health policy" (Parliamentary well-being group, 2014, pp. 33).

There might be multiple advantages in using SWB for public policy, and these are mainly liked to the broader prospective of SWB compared to HRQoL. In fact, health is only a contributing factor of the overall well-being of individuals (Stutzer and Frey, 2003), although being an extremely important one (Graham, 2008). Using a broader measure of benefit might help capturing aspects important to individuals that are at the moment excluded from the decision making process of resource allocation. For example, while being overweight or obese might not have an immediate health consequence, the problem might lead to substantial future healthcare costs. Issues like obesity, unless it is extremely severe, would most likely be poorly detected by a GPBM but well evidenced by a SWB measure. This is because, while the direct immediate impact of obesity on health might be limited, other aspects such as social relationships, social activities, social life and personal finance might be immediately affected by the condition. Developing a way to include SWB, or a greater component of SWB, in economic evaluations would therefore permit to widen the number of interventions that could be examined e.g. public health preventive policy against weight gain intervention. In addition, a wider measure of benefits would permit to increase the scope of economic evaluations by enabling cross sector resource allocation, therefore opening to the possibility of informing decision making on interventions that affect different budgets e.g. health and education. Finally, as SWB has been seen to be a good proxy of mental health (Mukuria and Brazier, 2013), and as mental health seems to be poorly detected by the majority of GPBMs (e.g. Richardson et al, 2015), SWB measures might help prioritizing areas of health which are neglected at the moment.

2.11 Conclusions

This chapter has presented four models of health to which most of the XXI century definitions relate. As some of these models refer to wellbeing e.g. WHO definition, it has explained what is wellbeing and it has discussed how both health and wellbeing are commonly associated to the term quality of life. It has clarified that a universally accepted definition of these terms and their bounderies does not exist. As definitions represent the framework for measuring what is intended to be measured, this lack of clarity on the bounderies of the definitions create problems when operationalizing the concept of health into measures of health. For these reasons, measures of health developed under the WHO framework and commonly referred to as HRQoL measures differ in the indicators they include i.e. dimensions and constructs despite aiming at measuring the same concept. The chapter has subsequently explained that outcomes are important for decision making, and that this is particularly true in the context of healthcare where market cannot operate unregulated. To inform healthcare decisions, a common approach is comparing costs and benefits of programs and interventions using economic evaluations. Three types of economic evaluations exist, and these differ in terms of the economic tradition i.e. welfarist or extra-welfarist to which they relate and the outcomes they use. Among them, CUA is the mostly used type of economic evaluation. It relies on the extra-welfarist tradition and measures health in terms of QALYs, a metric made of the length of life multiplied for the HRQoL of individuals in terms of utility. The chapter has subsequently described 3 common approaches for measuring HRQoL in terms of utility, and three techniques that can be used for this purpose. As HRQoL is not the only metric for which life years can be adjused, the chapter has presented SWB as an alternative. The next chapter focuses on GPBMs, a particular type of HRQoL measures based on utility.

Chapter 3
Generic preference based measures, their performance and solution to psychometric problems

3.1 Introduction

The previous chapter has introduced what are health and wellbeing, why economic evaluations should be used to inform resource allocation decisions in the health care sector, which approaches have been employed to derive utilities and which techniques can be used for this purpose. This chapter focuses on one of the approaches described, using GPBMs, as this is the most commonly used for obtaining health state utility values.

This chapter starts by introducing some of the key issues that need consideration when developing a new GPBM. On the one hand, this follows the discussion on the difficulty of operationalizing measures of health presented in Chapter 2, explaining pragmatically what choices measures developers face and why these might result in measures that differ in the dimensions and items they include. On the other hand, this introduces important aspects that will be useful later on in this thesis, such as why descriptive systems of GPBMs cannot include all possible dimensions of health, and what properties the dimensions of a GPBM should have. It then presents the five mostly used GPBMs and explains why consistency in decisions can only be ensured using one measure in all assessments. As a key characteristic of GPBMs is their appropriateness for the population for which they are used, the chapter summarizes the evidence on the validity and responsiveness of the five most commonly used GPBMs across all conditions or disease areas. Having found that all GPBMs report problems of validity and responsiveness in some conditions or disease areas, the chapter presents the solutions that are commonly employed when the reference GPBM lacks validity and responsiveness. One of these solutions, using bolt-ons, allows to maximize the cross-program comparability by ensuring consistency in the use of the core descriptive system of the reference GPBM and its valuation techniques. Having identified in bolt-ons a useful solution, the chapter investigates the bolt-on studies conducted to date through a scoping review. It finds that despite bolt-ons are a promessing area of research, they are currently identified and selected using different methods, and these have never been explored for this purpose. Moreover, it is not clear whether these are the only methods for identifying and selecting bolt-on dimensions. Having identified this gap, the chapter presents the aim of this thesis, which is examining methods that could be used to identify and select bolt-ons for GPBMs.

3.2 Generic preference based measures

GPBMs are measures of HRQoL commonly used to derive utility values for CUA. Similar to other health measures presented in chapter 2, also GPBMs differ in the elements of the continuum of health they include for the three aspects of health described in the WHO definition. For example, developers of the HUI 3 decided to adopt a within the skin approach, focusing on impairments and disability, and avoided the use of descriptions of social participation (Fenny, 2002; Fenny et al, 2002). By contrast, developers of the AQoL 8D decided to include domains for each category of the spectrum for the three aspects of health (Brazier et al, 2007).

Although GPBMs differ in their conceptualization of health, they share a common structure that is made of a descriptive system and a scoring system. The descriptive system describes health using a combination of dimensions, items and levels, generating a set of health states. The scoring system gives a utility value to each of these health states based on one of the three valuation techniques described in the previous chapter, or a mix of them. This utility value can be used to calculate QALYs for economic evaluations. Given these similarities, similar issues need to be considered when developing a new GPBM. These are described in the next section.

3.2.1 Issues around developing a GPBM

The first step required for the development of a new GPBM consists in identifying the dimensions (or constructs) that are relevant to describe health according to the conceptualization embraced by the developers. Identification of dimensions should rely on the use of interviews with patients or members of the public, where the importance of different constructs of health is examined (Cadman et al, 1986). However, dimensions have been historically identified based on expert opinion (Brazier et al, 2017).

After a set of dimensions has been identified, items for those dimensions need to be constructed. Constructing items refers to the process of developing a descriptor e.g. Mobility and labels e.g. I have no problems in walking about related to the construct of interest. Multiple items may be constructed for the same dimension. For example, the dimension pain can be measured through an item such as frequency of pain, or an item such as degree of pain.

When constructing items, attention should be paid in ensuring that these are easy to understand and to respond to (Fayers and Machin, 2007). A number of consideration have to be made for this purpose, among which avoiding double negatives, wording dimensions in a way that reflects common spoken language, avoiding ambiguous wording and presenting little variations in terms of dimensions wording i.e. consistency between dimensions (Fayers and Machin, 2007). While items should be easy to understand and respond to, these also have to be brief as they will need to undergo a valuation study. Hence, a trade-off might exist in the level of detail provided to ensure clarity and the need of keeping the descriptors and labels coincise (Brazier et al, 2017).

Another important element of GPBM development consists in deciding how many items per dimension to retain. In most GPBMs a dimension is described in terms of one or two items, although occasionally more have been employed (Brazier et al, 2007). Once again, different methods can be used to establish the relevant items per dimension, and these might include conventional psychometric criteria e.g. amount of missing data, responsiveness to change over two points in time, correlation with the relevant dimension etc., valuation tasks e.g. asking responders to rank the set of items by their importance, or Rasch analysis i.e. a mathematical technique that converts categorical data into a continuous latent scale using a logit model (Tesio, 2003). Before a final choice of item/s per dimension is made, an important aspect to consider is their relevance and face validity. Relevance (in a GPBM) refers to the ability of an item to tap on constructs that are important for peoples' HRQoL. Face validity refers to the ability of an item to reflect the intended construct in a clear and unambiguous way. This latter psychometric characteristic originates from a correct construction of the items, as described previously. Items relevance and face validity can be generally checked using qualitative techniques.

As items can describe health using different levels of severity, measures developers also need to decide the number of levels per item. Items levels should be easy to understand and ranked clearly. A method that has been proposed to determine the number of levels needed is asking responders to rate categories on a rating scale (Brazier et al, 2007). Alternatively, Rasch analysis can also be used to understand the relative position of item response choice (Brazier et al, 2007).

Once items have been developed for each of the dimensions of interest, developers of GPBM might have to consider reducing the number of dimensions (and items) for the new GPBM. This is because GPBMs need to be scored through a valuation study, where large descriptive systems have been seen to lead participants to adopt simplifying heuristics (e.g. Lloyd, 2003). Although there is no clear rule on the number of dimensions and items that a GPBM should include (GPBMs vary between 5 and 35 items), it has been suggested that individuals can process only between 5 to 9 pieces of information at a time (e.g. Dolan et al, 1996). This is however indicative, as the number of items that could be valued varies depending on the preference elicitation technique employed.

Different approaches have been proposed for selecting items and these comprise both qualitative and quantitative techniques. For example, interviews or focus groups with members of the general public can be employed to investigate what is the importance attributed to different constructs of health, selecting only those considered mostly relevant (Cadman et al, 1986). Alternatively, factor analysis might be a useful tool for reducing the number of dimensions or showing when dimensions are not independent (Young, 2005).

Finally, once a set of items for the dimension of interest has been constructed, tested in terms of relevance and face validity and reduced to be amenable to valuation, a valuation survey can be conducted. The valuation exercise is usually performed on members of the general public selecting a subset of the possible health states described by the GPBM. Values for the remaining health states are subsequently obtained through statistical modelling. The two valuation methods that received greatest favour for assigning utility values to GPBMs are the TTO and the SG. The rationale for this preference is that both these methods are choice based scaling techniques (Brazier et al, 2005; Dolan et al, 1996; Schulper 1996; Drummond et al, 2005).

A limited number of GPBMs dominate the literature (Richardson et al, 2015), and these are the EQ-5D, the SF-6D, the HUI3, the AQoL 8D and the 15D (Brazier et al, 2005). These are discussed in the next sections.

3.2.2 EQ-5D

The EQ-5D was developed by a multidisciplinary group of experts from five countries (EuroQol group, 1990). It is a five health dimensions measure, with dimensions being selected based on experts' opinion. These are: mobility, self-care, usual activities, pain/discomfort, and anxiety and depression; each dimension is given one of three levels of severity, which results in 243 possible unique health states. The most widely used scoring algorithm was obtained through a valuation study conducted by the UK Measurement and Valuation of Health (MVH) group at York. Values were obtained asking respondents to elicit 43 health states using VAS and TTO. Regressions were employed to model the data (Brazier et al, 2007). However, value sets for other countries exist, and more specifically for: Belgium, Denmark, Finland, Germany, Japan, New Zeeland, Slovenia, Spain, United States and Zimbabwe (EuroQol Group, 2009). These were obtained using both TTO and VAS and have been found to substantially differ from the UK ones. For this reason, the use of country specific values is recommended if possible (Brazier et al, 2007). In 2005, a task force was established within the EuroQol group to investigate ways for improving the instrument sensitivity to changes in health (Herdman et al, 2011), which led to the development of the EQ-5D-5L. The new version of the instrument maintains the same number of dimensions of the EQ-5D, but the levels for each dimension were expanded to 5 levels (5L). The version has been translated in more than 126 languages and a cross walk value set is available from Van Hout et al (2012). The new value set for England was released recently, and is available in Devlin et al (2016).

3.2.3 SF-6D

The SF-6D was developed by a team at the University of Sheffield (Brazier et al, 1998; Brazier et al, 2002). It is a six-dimension classification system that takes advantage of the mostly used health state classification system in the world, the Short Form 36 (SF-36) (Brazier et al, 2007). The six dimensions were identified from the 36 items of the SF-6D with the objective of generating a descriptive system amenable to valuation. This was done using a set of criteria, among which avoiding redundancy between items, giving preference to negative items and using people preferences when available (Brazier et al, 1998). Dimensions are: physical functioning, role limitation, social functioning, pain, energy, mental health; each dimension has either four or six severity levels, which results in 18,000 possible unique health states for the version obtained from the SF-36 (Brazier et al, 2002) and 7500 possible unique health states for the version derived from the Short Form 12 (Brazier and Roberts, 2004). The original valuation study was conducted using the SG for 249 states and modelling data through a random effect regression model (Brazier et al, 2002; Brazier and Roberts, 2004). A new algorithm has also been developed using a non-parametric Bayesian approach (Kharroubi et al, 2005).

3.2.4 HUI 3

The HUI3 was developed by Feeny et al (2002) as an evolution from the previous measures Health utility index 1 (Torrance et al, 1982) and Health utility index 2 (HUI 2). (Torrance et al, 1996). In addition to the 6 dimensions of the HUI 2, the HUI 3 includes also vision and hearing. This addition was made based on measure developers' judgment in order to increase the measure's sensistvity. This resulted in an eight-health dimensions instrument that uses the same 15 item self-completed questionnaire of the Health utility index 2. Dimensions are: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain; each dimension has five or six severity levels, which results in 972,000 possible unique health states. The valuation survey has been conducted in Canada and France using VAS and SG and asking responders to value first a single dimensional state, then a set of corner state with one dimension at its worst and all the other at their best, and then a multiattribute utility state. (LeGales et al, 2002).

3.2.5 15D

The 15D was developed by Sintonen (1994) selecting dimensions based on experts opinion and from a review of Finnish policy documents. It is a fifteen-dimension instrument. Dimensions are: mobility, vision, hearing, breathing, sleeping, eating, speech, excretion, usual activities, mental function, discomfort and symptoms, depression, distress, vitality, sexual activity; each dimension has either four or five severity levels, resulting in billions of unique health states. The valuation survey used a variant of the VAS.

3.2.6 AQoL 8D

The AQoL 8D was developed by Hawthorne, Richardson and colleagues at the Universities of Melbourne and Monash. It is an eight health-dimensions instrument from a 35 items questionnaire identified through a review of the literature, focus groups and experts' opinion. Dimensions are: independent living, happiness, mental health, coping, relationships, self worth, pain, senses; each dimensions has a variable number of question, each of which has four, five or six severity severity levels, which results in 2.4 x 1023 possible unique health states. The valuation of health states followed a two-step approach. The first step used a multiplicative model to combine items into dimension. A second step adjusted for the double counting due to the absence of an orthogonal design. Values were obtained using TTO and VAS (Richardson et al, 2011).

3.2.7 Summary

In summary, different methods can be used to identify the dimensions of a GPBM, construct its items, select between the identified dimensions and items and determine the number of levels through which the item should be described. However, from an empirical point of view, development of most GPBMs has havily relied on experts' opinion. This has resulted in the 5 mostly used GPBMs covering different health dimensions, including different items, presenting different number of levels per item and differing in the valuation methods they employed.

3.3 An overview of reviews of the performance of the five mostly used generic preference based measures

3.3.1 Introduction

As previously mentioned, GPBMs differ in terms of the dimensions they cover, the items they use, the number of levels for their items and the elicitation techniques through which the health states they describe are valued. It is therefore not surprising that they generate substantially different utility values (Nord et al, 1993).

Differences in GPBMs mean score have been seen to be typically between 0.03 and 0.05 between the SF-6D, EQ-5D, HUI 3 and the AQoL, although differences with the 15 D tend to be higher (e.g. Brazier et al, 2004; O'Brien et al, 2004; Richardson et al, 2015 a). Whilst a variation of 0.03 might seem negligible, this translates into a 3% change in the risk of death in the SG, and into a variation of 3.5 months on a 10 years time period in the TTO (Brazier et al, 2007). An example of the size of this impact on economic evaluation has been provided in a study of Bansback and colleagues (2007), where utilities obtained from the EQ-5D and the HUI 3 were used to assess a treatment for macular degeneration. The study found that using the HUI 3 the cost per QALYs of the intervention arm resulted of approximately 20,000 pounds/ QALY, while using the EQ-5D of approximately 140,000 pounds/ QALY.

It has been suggested that both the valuation set used by the measures and the number of levels per dimension are possible contributing factors for these variations (Brazier et al, 2007). However, a third element, the ability of the descriptive system to tap on dimensions relevant for the condition under investigation has been shown to be the main cause of differences in a study of Richardson et al (2015). The study adopted a regression based approach to investigate the extent of, and the reason for differences between the utilities predicted by the EQ-5D-5L, SF-6D, HUI 3, 15D and AQoL 8D, by comparing the role of scale effects i.e. differences due to number of levels, micro-utility effects i.e. scale adjusted differences due to the utility formula and the content of the instrument in explaining discrepancies between predicted utilities. The study found that, overall, 66% of the differences was attributable to the descriptive system of the GPBMs, 30.3% to scale effects, and only the remaining 3.7% to micro-utility effects. The importance of scale effect, micro-utility and descriptive systems in explaining overall variation in GPBMs values varied substantially when different pairs of measures where compared, with scale effects being more important than the content of the descriptive system in some cases. Nevertheless, the impact of the descriptive system was always significant, accounting between the minimum of 24.7% of the overall variation in the comparison between the 15D and AQoL 8D, and the maximum of 101.6% of the overall variation in the comparison between the HUI3 and AQoL 8D.If GPBMs generate different valuesthe question arises whether these are appropriate for the group of patients being examined in the evaluation and able to detect meaningful changes.

There are numerous criteria for assessing measure’s appropriateness, and these are their practicality i.e. the acceptability of an instrument in the population in which it is going to be employed, their reliability i.e. the stability of an instrument over subsequent administrations in presence of stable health, their construct validity i.e. the ability of an instrument to measure what it is intended to measure and their responsiveness i.e. the ability of an instrument to detect changes in health when these have occurred (Walters, 2009; McDowell, 1987). Although all these criteria are important, the construct validity and responsiveness of a measure are the most fundamental ones, as it makes little sense to have a GPBM that is practical and reliable but unable to measure the right concepts and detect changes in patients when these have occurred (Brazier et al, 2007).

There is a growing body of literature investigating the construct validity and responsiveness of GPBMs in either a specific population or for a specific GPBM. This evidence has been increasingly summarized in a number of systematic reviews, but there has been no overview of this evidence across GPBMs. Hence, this evidence is disaggregated, making it difficult to draw conclusions regarding the overall performance of GPBMs.

This section addresses the gap by reporting an overview on the construct validity and responsiveness of five GPBMs, including the coverage and nature of the evidence in different conditions, based on existing systematic reviews.

3.3.2 Methods

An overview of reviews was undertaken. Consistent with the Cochrane collaboration guidelines (Higgins and Green, 2011) all phases of this study were planned and summarized in an overview protocol. Formal guidance on reporting of overviews of systematic reviews does not exist but whenever possible, the 27 item checklist covering important information needed in reporting systematic reviews and meta-analysis of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) were followed (Moher et al, 2009).

Search strategy and study identification

A search strategy was developed to identify reviews on the validity and responsiveness of the five most commonly used GPBMs for adults, across all disease classes. The search combined free text and controlled vocabulary words, among which “quality of life”, “patient reported outcome”, “preference based instrument”, “psychometric characteristic”, “EQ-5D”, “SF-6D”, “HUI 3”, “AQoL” and “15D”, all with spelling variations, acronyms and related terms (Appendix Chapter 3 - Table I).

A highly sensitive searching filter for reviews and meta-analysis developed by the information services team of the Canadian Agency for Drugs and Technologies in Health was used to refine the search (CADTH, 2014), which was not limited by date or language restrictions. Medline, Embase, Cochrane Library and Scharr HUD electronic databases were investigated. In addition, references of the included reviews were screened and a complementary search on Google Scholar was performed.

Identified citations (both published and grey literature) were assessed against the following set of pre-defined eligibility criteria. Reports were eligible for inclusion if they were reviews, they examined the construct validity or responsiveness of at least one GPBMs, their main focus was on adult population (defined as ≥18 years old) and they summarized results reporting information at the study level (either in the text, tables or appendix of the reviews). Reviews were excluded if they reported results only in aggregate form, if they only examined reliability or face validity (i.e. not construct validity or responsiveness), if they tested translations of the GPBMs, if they were not in English or if they were only poster presentations.

Quality Assessment of the reviews

Quality has been assessed using a modified eight question version of the AMSTAR checklist for systematic reviews (Shea et al, 2007) with items weighted for their importance for GPBM research (See Appendixes Chapter 3 - Table II). Weighting was agreed after discussion between the PhD student and the supervisors John Brazier and Clara Mukuria. Questions on the “comprehensiveness of the literature search”, the “presence of a quality assessment tool” and the “use of quality scores to formulate conclusions” were assigned two points as they were considered essential for the correct identification and assessment of quality of studies included in reviews. “Characteristics of the included studies” was assigned 1.5 points, as these might significantly impact on the results. The “presence of duplicate data selection and extraction” and “double blinding” (although rarely used in systematic reviews of psychometric evidence) were assigned a score of one since they strengthen the reliability of the selection process. Questions on providing an *a priori* design, which minimizes the chance of results being changed once searches have being completed, a list of included studies and conflicts of interest were given a weight of 0.5 as these were considered to have less of an impact on reviews of psychometric studies. Questions in AMSTAR on the “methods used to combine findings”, the “likelihood of publication bias” and the “status of publication used as an inclusion criterion” were excluded because they were considered irrelevant for systematic reviews of measures’ psychometric performance.

The resulting checklist has a minimum score of 0 and a maximum score of 10. As a way to categorize the quality of systematic reviews, arbitrary cut-offs were assigned, considering reviews of excellent quality if they received a score ≥7.5, of good quality if they received a score ≥5 and of poor quality with a score 5<. Scores for both the original and the modified checklists are provided in the Appendices Chapter 3 - Table III.

Data extraction

A customized extraction template was designed and piloted on 5 reviews. Information on reviews characteristics (e.g. review objectives, number of studies included, disease classes investigated, condition examined) and details of the psychometric assessments undertaken were extracted. In the case of a review published in several places, then the article with the most up to date data was used supplemented by additional evidence contained in the other sources. When different reviews included the same study, the most complete data for that study were extracted, supplemented by the evidence contained in the other review and presented in the results for only one of the two reviews to avoid double counting of studies.

Assessment of findings

Validity

The validity of an instrument should ideally be assessed by comparing it to a gold standard measure of the construct of interest. Where a gold standard or criterion does not exist, psychometricians use indirect indicators of validity (Brazier and Deverill, 1999). One indicator is the ability of an instrument to distinguish between groups known or thought to differ in the trait or behavior, such as defining groups by severity of condition or patients vs. general population. Assessment of whether or not known group validity is evident can then be based on whether those with poorer health also have lower utility scores, using appropriate tests to assess whether these differences are statistically significant (e.g. t-tests) and important in magnitude (e.g. using standardized effect sizes (SES), which is the difference in the scores divided by the pooled standard deviation).

Another indicator is convergent validity, which examines the extent to which two measures of the same or similar concept agree with each other, for example by using correlations. The magnitude of the correlation is used to judge the extent to which GPBMs are related to the comparison measure.

Responsiveness

Responsiveness focuses on a measures ability to reflect changes that have occurred in health (Wilkin et al, 1992; Brazier and Deverill, 1999), such as by comparing patients before and after a successful treatment. Change is usually assessed based on whether differences in utility scores are statistically significant and their standardized magnitudes large, using standardized effect sizes (SES) or standardized response means (SRMs) (i.e. the change divided by the change standard deviation).

Criteria for psychometric assessment

Criteria are required to judge whether measures meet the psychometric properties being assessed. Cohen's criteria have been used in this overview (Cohen, 1977). Correlations are very strong if >0.6; strong between 0.5-0.6; moderate between 0.49-0.3; and weak if 0.29≤ (Cohen, 1977). Moderate to very strong correlations were taken as an indicator of convergent validity. SES and SRMs were judged as large if they were ≥ 0.80; moderate between 0.50 and 0.79; and small between 0.2 and 0.49 (Cohen, 1977). Moderate to large SES and SRMs were taken as a sign of construct validity or responsiveness. Statistical significance was also considered as evidence to support of known group validity and responsiveness. These criteria only provide indicative guidance on the psychometric characteristic of an instrument. Judgements must also be used based on the quality of studies included and the characteristics of the indirect indicators that are used.

Reporting

Evidence is presented in summary tables by measure and condition and reviewed by narrative synthesis. In the summary tables, symbols are used to identify where evidence supports validity or responsiveness (✓), suggests poor validity or responsiveness (✗), is mixed (±), which indicates some supporting evidence and some against, inconclusive (/), when evidence is lacking e.g. data too sparse, or NR when the measure is not reported in the review. Conditions are grouped using the international classification of diseases (ICD), which categorises all existing conditions into 22 classes (WHO, 2010). Each ICD class includes a variable number of conditions between 45 and 99. AQoL 8D and 15D results are only presented in the text due to the limited evidence found.

3.3.3 Results

A total of 2216 potentially relevant articles were identified after removing duplicates. Title and abstract screening excluded 1661 and 465 records respectively, and full text screening excluded an additional 63. Online search and reference screening found 3 reviews that had not been detected by database searches. Consequently, 30 reviews were included (Bansback et al, 2008; Brazier et al, 2014; Castelino et al, 2013; Ching et al, 2003; Davis and Wailoo, 2013; Derrett et al, 2009; Devine et al, 2011; Dyer et al, 2010; Haywood et al, 2005; Hill et al, 2010; Holloway et al, 2014; Hounsome et al, 2011; Janssen et al, 2011; Kuspinar and Mayo, 2014; Linder et al, 2003; Longworth et al, 2014; Papaioannou et al, 2013; Papaioannou et al, 2011; Peasgood et al, 2012; Petrillo et al, 2011; Pickard et al, 2007; Pickard et al, 2008; Sanghera et al, 2013; Speight et al, 2009; Szende et al, 2003; Tosh et al, 2012; Whitehurst et al, 2012; Wu et al, 2013; Yang et al, 2013; Yang et al, 2014). Figure 4.1 summarizes the selection process. A list of included and excluded reviews is provided in Appendices Chapter 3 - Table IV and Appendices Chapter 3 - Table V.

Characteristics of the included reviews

The number of studies included in the reviews varied significantly[1], from five (DeVine et al, 2011) to 122 (Haywood et al, 2005). Most reviews included a mix of randomized clinical trials (RCTs), cross sectional, cohort and longitudinal studies, or a mix of other experimental and/or observational designs, apart from Devine et al (2011) which focused on longitudinal studies and Holloway et al (2014) which focused on RCTs. One review by Bansback et al (2008) included only economic evaluations. Table 1 summarizes the main characteristics of the included reviews.

Quality of included reviews

Two reviews (Kuspinar and Mayo, 2014; Linder et al, 2003) received an assessment of excellent quality and 14 of good quality (Brazier et al, 2014; Davis and Wailoo, 2013; Jansenn et al, 2011; Longworth et al, 2014; Papaioannou et al, 2013; Papaioannou et al, 2011; Peasgood et al, 2012; Pickard et al, 2007; Tosh et al, 2012; Whitehurst et al, 2012; Wu et al, 2013; Yang et al, 2013; Yang et al, 2014; Szende et al, 2003,). The remaining 14 reviews received a poor quality assessment (Bansback et al, 2008; Castelino et al, 2013; Ching et al, 2003; Derrett et al, 2009; Devine et al, 2011; Dyer et al, 2010; Haywood et al, 2005; Hill et al, 2010; Holloway et al, 2014; Hounsome et al, 2011; Petrillo et al, 2011; Pickard et al, 2008; Speight et al, 2009; Sanghera et al, 2013). The main reason for poor quality was that reviews did not assess the quality of the included papers themselves and consequently, did not consider scientific quality appropriately in drawing conclusions. Five reviews received an AMSTAR modified score below 3, with four of them reporting a literature search that was not considered comprehensive (i.e. terms were not derived paying attention to synonyms, acronyms and related terms for the building blocks of the research question) (Castelino et al, 2013; Ching et al, 2003; Petrillo et al, 2011;

[1] Not all the studies included in the reviews were relevant to the research question investigated in this overview. However, this overview draws on more than hundred and eighty studies included in the thirty reviews.

Speight et al, 2009) and none of these performed a double blind study selection (Castelino et al, 2013; Ching et al, 2003; Petrillo et al, 2011; Sanghera et al, 2013; Speight et al, 2009).

Breadth and depth of the evidence

Twenty-nine reviews reported information for the EQ-5D, twelve for the SF-6D, eight for the HUI3, two for the 15D and three for the AQoL 8D.

EQ-5D psychometric characteristics were presented for conditions across 16 of the 22 ICD classes of disease codes (Table 3.2). Two reviews reported EQ-5D characteristics in a class not specified (i.e. aesthetic surgery in Ching et al (2003) and older population in Haywood et al (2005). SF-6D psychometric performance was reported for conditions related to 9 classes of disease, HUI3 to 7 classes, and 15 D and AQoL only to 2 classes of disease. Very few conditions were covered within each ICD class. For example, of the 99 diseases included in the ICD class diseases of the respiratory system only asthma and COPD were investigated, and of the 90 diseases included in the ICD class endocrine, nutritional and metabolic diseases only type 2 diabetes.

The amount of evidence in relation to the psychometric assessment of validity and responsiveness within conditions varied substantially, with some reviews reporting multiple psychometric analysis results and others focusing on a single type of assessment. Overall there was much less evidence available for measures other than the EQ-5D.

Type of evidence

Known groups testing

Of the 180 studies included in the systematic reviews that reported known groups validity (more detailed results available in Appendices Chapter 3 – Table VI), 77 used comparisons based on severity traits although two studies did not use all the potential severity levels (Petrillo et al, 2011; Dyer et al, 2010). For the other studies comparisons were based on patients versus general population (44 studies), different types of diseases or disorders (15 studies), groups defined by an HRQoL instrument (7 studies), numbers of disease/disorders (4 studies) and patients with or without complications (3 studies). Comparisons were also based on other groups such as discharged and not discharged patients (21 studies). Nine studies used groups that

were considered inappropriate for testing GPBMs validity, like age, education, different country cohorts and income. Most studies assessed known groups based on utility scores but seven reviews (Davis and Wailoo, 2013; Kuspinar and Mayo, 2014; Longworth et al, 2014; Tosh et al, 2012; Yang et al, 2014; Pickard et al, 2008; Szende et al, 2003) reported results for the unscored dimensions of the instruments.

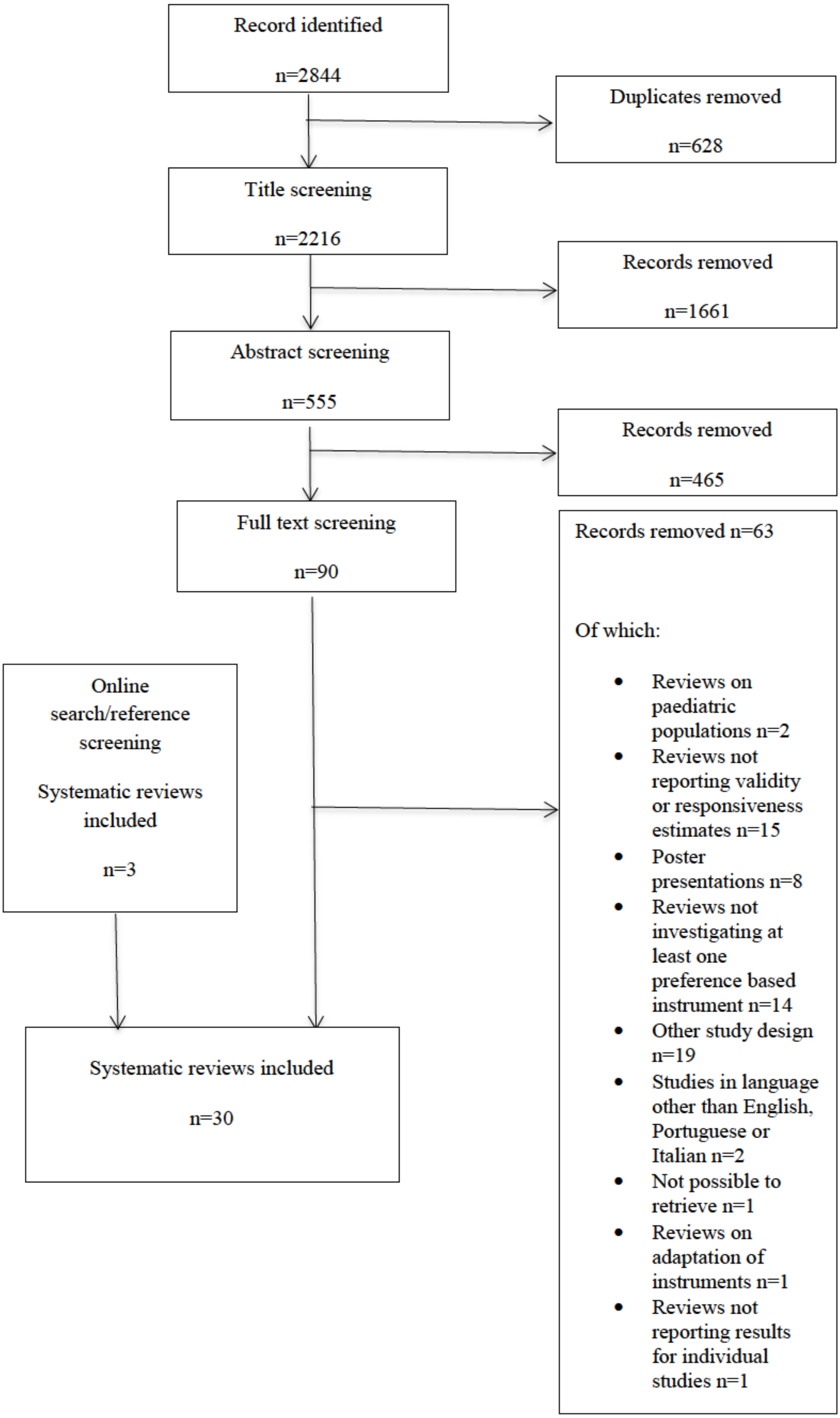
Convergent validity

Correlations with other measures were reported in 135 studies (more detailed results available in Appendices Chapter 3 – Table VI), 38 of which used a non-preference-based HRQoL measure, 32 used a direct utility measure (e.g. TTO), 27 used a symptom or severity measure, 18 used a functional status measure, 9 used another GPBM and 14 did not specify the measure used.

Responsiveness

Reviews reported 172 studies on GPBMs responsiveness (more detailed results available in Appendices Chapter 3 – Table VIII) most of which (n=124) were based on comparing patients before and after a successful treatment, with 112 of these reporting statistically significant differences, 8 reporting SESs, 2 reporting SRMs and 2 not reporting the method employed. Comparisons were also based on patient groups receiving different treatments (n=38; 32 reporting statistical significance and 6 reporting SESs), and patients reporting an improved health state (n=6; 3 reporting SESs and 3 reporting SRMs). 4 did not specify the groups used but reported SRMs.

Figure 3.1 – Flow chart of the selection process



Instruments performance by condition

The overwhelming majority of evidence in type 1 (Janssen et al, 2011) and 2 (Janssen et al, 2011; Speight et al, 2009) diabetes mellitus showed that EQ-5D possessed good discrimination between severity groups, correlated moderately to strongly with other HRQoL instruments and reported changes consistent with expectations after patients' treatment. Little evidence was found for the SF-6D, and this was mixed.

The review on diseases of the skin and subcutaneous tissues (among which psoriasis, acne, eczema and legs ulcers) (Yang et al, 2014) presented results supporting EQ-5D validity and responsiveness, with only 2 out of 27 studies reporting evidence against the measure's validity, which were weak correlations and lower SRMs for EQ-5D compared to other measures.

Two systematic reviews investigated COPD and Asthma (Petrillo et al, 2011; Pickard et al, 2007), suggesting that the EQ-5D is generally valid based on known group comparisons of severity and patients/general population groups and correlations between the EQ-5D and non-preference based HRQoL measures. Results for responsiveness were mixed, with two studies reporting weak SRMs of the measure, one study strong SRMs and four showing changes in the expected direction using SESs and statistical significance. The only comparative study across GPBMs reported poor correlations between EQ-5D and SF-6D.

One review each investigated the performance of the EQ-5D in urinary incontinence (Davis and Wailoo, 2013) and HIV (Wu et al, 2013). There was evidence of validity and responsiveness in urinary incontinence (Davis and Wailoo, 2013) with five studies supporting discriminative validity based on severity levels and type of urinary incontinence, seven reporting moderate to strong correlations with HRQoL and symptom and severity measures, and five showing differences in health status from baseline to follow up and between treatment arms. Two studies reported mixed results, one showing that the EQ-5D distinguished between some types of urinary incontinences but not others and the other that the EQ-5D detected treatment differences only for some groups of patients, where other measures registered changes for all treatment groups. Two studies had inconclusive results for convergent validity as they did not specify the strength of correlations between measures. One study reported results for other GPBMs, supporting SF-6D, 15D and AQoL known group validity based on the assessment of severity traits. In HIV (Wu et al, 2013) responsiveness of the EQ-5D was weak, showing generally small before and after treatment SESs in presence of moderate or large ESs for the comparator measures. The only study investigating construct validity reported a good ability of the measure to discriminate between known groups.

The EQ-5D appeared generally valid and responsive in a number of cancers (Longworth et al, 2014; Pickard et al, 2007) (among which lung, breast, cervical, colon, kidney, liver cancer and leukemia) although limitations were found in some studies. Twenty-five of the 31 studies examining known groups differences showed that EQ-5D distinguished between cancer severities, patients/general population and groups with different types of cancer; 12 of the 17 studies examining convergent validity reported moderate to strong correlations with direct utility measures, HRQoL measures and functional status measures; and 29 of 43 studies examining responsiveness showed that the measure detected changes between treatment arms and from baseline to follow up that were consistent with those of comparator measures. A significant amount of evidence supported HUI3 psychometric characteristics (Longworth et al, 2014; Pickard et al, 2007) with 8 studies out of 11 showing good discriminative ability in distinguishing between severity levels, type of cancer and cancer patients/ general population, 4 studies out of 7 reporting good convergence with functional status measures and 8 studies out of 10 a good ability to detect changes from baseline and between treatment arms. Only two studies reported information for the SF-6D. In one, the measure was not able to detect differences between cancer patients and the general population. In another, the measure correlated appropriately with a cancer HRQoL questionnaire. Very few comparative studies were reported between the investigated GPBMs, and these do not clarify which performs better.

The EQ-5D showed a mixed performance in cardiovascular diseases (Dyer et al, 2010) (among which coronary hearth disease, cerebrovascular disease, hypertension and hearth failure). Although many studies supported the instrument convergent validity with other GPBMs, HRQoL measures and functional status measures and its ability to distinguish known groups based on severities of the conditions and type of conditions, two studies showed poor correlations with a HRQoL measures, three problems in distinguishing between patients and the general population, eight failed to detect statistically significant changes at follow-up and one failed to show differences between treatment arms. Three comparative studies were reported between the EQ-5D and SF-6D, the EQ-5D and HUI3, and the EQ-5D, SF-6D and HUI3. In two of them, correlation between the EQ-5D and SF-6D, and between the EQ-5D, HUI3 and SF 36 were generally poor. The third comparative study presented moderate to strong correlations between the three instruments.

The EQ-5D performance in visual disorders (among which macular degeneration, glaucoma, conjunctivitis, diabetic retinopathy and others) (Tosh et al, 2012) was generally mixed. Known groups showed generally poor or mixed validity using severity groups, and generally good validity using patients versus general population groups. Mixed evidence was also reported for convergent validity, with the instrument correlating moderately to strongly with clinical measures only in four of the nine studies that investigated the property. There was mixed and limited evidence for the EQ-5D responsiveness, with one study reporting in support, one against and one mixed evidence for the measure characteristic. All these studies used tests of statistical significance before and after treatment. The HUI3 appeared to be valid although the evidence was limited. Two studies reported a good ability of the measure to distinguish known groups based on the severity of the condition and on patients/general population. Another study reported moderate to strong correlations with functional status measures. A fourth study showed that the HUI3 was able to detect statistically significant changes between treatment arms (Tosh et al, 2012). Only two studies reported on the SF-6D characteristics and these showed the measured performed better than the EQ-5D (Tosh et al, 2012).

The EQ-5D performance has been reviewed in only one condition of the nervous system (Kuspinar and Mayo, 2014), multiple sclerosis, with three studies supporting the instrument convergent validity and three reporting weak to moderate correlations with other HRQoL measures. Substantial evidence against the instrument's ability to distinguish between severity groups was found, with two studies reporting the measure distinguished only between some severity levels but not others (mixed evidence), and two showing the measure was not able to detect health status differences in any of the severity levels. Evidence for the SF-6D, HUI3 and AQoL was limited but in support of the measure performance (Kuspinar and Mayo, 2014), with two studies reporting moderate to strong convergence of the SF-6D with HRQoL measures, two good discriminative ability of the HUI3 between severity groups and strong correlations of the measure with other HRQoL instruments and two good discriminative ability of the AQoL 8D, with the assessment being based on severity levels.

The EQ-5D performance in hearing impairments (Yang et al, 2013) was poor, with only two studies out of the seven supporting validity and responsiveness, one reporting moderate to strong correlations with other GPBMs and the other statistically significant changes of score before and after treatment. The HUI3 showed a better performance, with all known groups assessments but one in favour of the instrument validity (based on severity traits) and most of the responsiveness tests showing an ability to detect changes in health status before and after treatment (Yang et al, 2013). Although few comparative studies were found, all these suggested that the HUI3 performs better than EQ5D in hearing impairment.

Five reviews investigated the performance of the EQ-5D in mental health (Brazier et al, 2014; Papaioannou et al, 2013; Papaioannou et al, 2011; Peasgood et al, 2012; Hounsome et al, 2011), and all but the one on depression and anxiety showed that the instrument suffered from problems. Three studies showed low correlations between the EQ-5D and HRQoL measures in dementia, four low correlations between the EQ-5D and the time trade off, standard gamble and symptom specific measures in schizophrenia, two low correlations between the EQ-5D and other measures (not specified) in bipolar disorder and two low to moderate correlations between the EQ-5D and symptom and severity measures in personality disorders. Evidence against the measure validity was found also for known groups in personality disorders and bipolar disorder, with one study showing poor discrimination between groups based on different types of personality disorders and another poor discrimination between severity levels of bipolar disorder. Convergent validity, known groups and responsiveness results for the SF-6D and HUI3 supported the instruments psychometric characteristics, with the exception of an SF-6D known group test that showed mixed results in depression (discriminating only between some groups but not others) (Peasgood et al, 2012), although the evidence base was smaller.

Four systematic reviews reported evidence on EQ-5D and SF-6D psychometric characteristics in musculoskeletal diseases (Whitehurst et al, 2012; DeVine et al, 2011; Hill et al, 2010; Bansback et al, 2008). One study reported good convergence for the EQ-5D with another HRQoL measure in rheumatoid arthritis, while another inconclusive results in chronic low back pain, with data being too sparse to assess correlations. The SF-6D was seen to have moderate to strong convergence with an HRQoL measure in rheumatoid arthritis, but mixed known group results in spinal cord injuries, with three studies supporting the instrument discriminative ability and four reporting against it (Whitehurst et al, 2012).

Evidence for the other disease ICD classes was very sparsely investigated, including hematological, gynecological and autoimmune diseases, and diseases of the nose. Three reviews investigated injuries, aesthetic surgery and older populations, but evidence was extremely limited, although the few studies available were generally in support of GPBMs psychometric characteristics (Davis and Wailoo, 2013; Pickard et al, 2007, Whitehurst et al, 2012; Devine et al, 2011; Haywood et al, 2005; Bansback et al, 2008; Ching et al, 2003; Holloway et al, 2014).

3.3.4 Discussion

The aim of this overview of reviews was to summarize the evidence on the validity and responsiveness of five GPBMs, the EQ-5D, SF-6D, HUI3, AQoL 8D and 15D in terms of the size, quality and nature of the evidence across different conditions, and to determine whether it is possible to draw conclusions about their relative performance. A systematic review of reviews was undertaken that yielded 30 systematic reviews, which included more than 180 studies.

Size and coverage of the evidence

The body of evidence was heavily skewed towards EQ-5D, with significantly fewer systematic reviews investigating HUI3 and SF-6D, and almost none 15D and AQoL 8D. Furthermore, the number of conditions covered was limited even for EQ-5D. There were also limitations in the psychometric assessment that was reported. For example, some studies only reported convergent validity, or they limited assessments to comparisons with only one other indicator. This limits the conclusions that can be drawn from the evidence across the different GPBMs and within some conditions.

Quality, nature and reporting of evidence

Many of the reviews received an AMSTAR modified score of poor quality mainly because they did not assess the quality of the studies they included and consequently the impact of this on their synthesized results. In the presence of discordant results between studies, quality assessment can help in the interpretation and synthesis of evidence, for example by giving greater weight to more robust reports.

Reviews reported different types of evidence for each of the two indicators of validity and responsiveness such as known groups being defined by severity, number of diseases/disorders or simply patient versus general population, and treating them as equally informative. Although this is common practise in empirical studies, some tests should be considered more appropriate than others. For example, trait severity may be considered more informative than the trait number of disease/ disorders, since the ranking of preferences might be ambiguous in the latter case e.g. one severe condition might be worse than two mild ones. Comparing patient and general population scores is likely to be very crude. Furthermore, the tests often rely on clinical assessments that may not reflect the HRQoL of patients or preferences for the states. These aspects need more carefully consideration in the phase of reviews analysis and synthesis, as well as for the design of primary studies.

A number of concerns exist in the way in which evidence was reported by the included reviews. Few reviews stated with clarity which thresholds were adopted in analyzing and summarizing results, making the interpretation of the definitions “strong”, “moderate” or “weak” more difficult. It was also frequent to find outcomes defined as “significant”, and doubts remained on whether “significant” meant statistically significant or size or both. Some known groups tests based on severity of the condition reported only part of the range of possible severity levels. This significantly weakens the value of the evidence produced.

Performance of instruments

Despite the lack of evidence and standardisation across the reviews or studies included in psychometric assessment, some broad conclusions can be drawn from this overview of reviews. Where evidence is available, it often supports the GPBMs performance. EQ-5D appeared valid and responsive in conditions of the skin, respiratory, genitourinary, endocrine, nutritional and metabolic diseases, and for the majority of cancers where there is evidence; SF-6D was found to be valid and responsive in mental health and in diseases of the eye, the nervous and the genitourinary systems; HUI3 showed good validity and responsiveness in cancer, diseases of the eye, the ear, the nervous systems and mental health; AQoL 8D presented good psychometric characteristics in the musculoskeletal and genitourinary conditions and 15D in genitourinary, diabetes, nutritional and metabolic diseases. However, any attempt to compare the instruments is limited by a lack of head to head comparisons and the little evidence available on all GPBMs except EQ-5D.

There was also evidence of lack of appropriateness of GPBMs in some conditions. EQ-5D was found to perform poorly in hearing impairments, multiple sclerosis, personality disorders, schizophrenia and dementia, and reported mixed results in visual disorders, cardiovascular diseases and in some cancers. SF-6D showed inconsistencies in its ability to converge with other measures in cardiovascular and respiratory diseases and to discriminate between groups in neoplasms, while HUI3 reported mixed results for some subpopulations of neoplasms.

Most of the evidence that was used in the reviews relied on studies that used existing datasets, but this provides limited answers when investigating GPBMs comparative performance and it highlights the importance of designing *bespoke* comparative studies for this purpose. There are a few examples of these, including two large and two smaller studies where five instruments were investigated (Richardson et al, 2014 a), and the more recent Multi Instrument Comparison (MIC) project (Richardson et al, 2014 b), that compared a number of GPBMs and other measures across different conditions. This evidence, which was not integrated in the reviews, shows that convergence between GPBMs is generally moderate to large, but that the mean change is driven by the different constructs covered by the their descriptive systems, with some measures being dominated by physical functioning dimensions like EQ-5D and others more by psychosocial dimensions like AQoL 8D. This might serve as an explanation for the lack of validity and responsiveness noticed in some disease areas for the investigated GPBMs.

Limitations of the overview

This overview of reviews has some important limitations. Psychometric properties of GPBMs in some conditions may have been missed because of the lack of a systematic review for those conditions. In addition, this overview has been limited by the poor reporting of some reviews /studies. This overview focused on the five most widely used generic GPBMs, but there are other methods for obtaining health state utility values which were not covered, such as condition specific preference based measures, bespoke vignettes or direct valuations of patients' health states. These alternatives may provide an important source of evidence, particularly where the existing generic measure do not appear to perform well.

Conclusions

Whenever evidence is available this often supports GPBMs performance. However, all GPBMs appear to have problems of validity and responsiveness in some conditions or disease areas. Moreover, psychometric evidence breadth and depth is inconsistent between ICD disease classes, conditions, instruments and type of assessment. Indeed, there is often no evidence at all, or what is available is severely limited in nature and quality, and rarely enables direct comparisons across measures. This highlights the need for more rigorous reporting of GPBMs psychometric studies and reviews. It also highlights the need for large comparative studies designed to test instruments performance. As previously mentioned, few of these studies exist, and these show that the mean differences between GPBMs is mostly driven by the different constructs covered by their descriptive systems. This might serve as an explanation for the lack of validity and responsiveness noticed in some disease areas for the investigated GPBMs.

3.4 Rationale for bolt-ons

As seen in the previous section GPBMs might present problems of validity or responsiveness in some conditions or disease areas. Two possible explanations may be given for GPBMs lack of validity and responsiveness. The first is that the number of levels in which the health dimensions are described is not sufficient to capture changes in patients' HRQoL for the investigated disease. The second is that important aspects of the HRQoL are not appropriately captured by the measure's descriptive system (Longworth et al, 2014). When the first problem arises an effective solution is increasing the number of levels for those dimensions lacking sensitivity, as it has been done with the development of the EQ-5D-5L (Herdman et al, 2011). When the second problem emerges, a different GPBM, a condition specific preference based measure or direct valuation of patients' health need to be used (Brazier et al, 2017).

Alternatives for addressing the lack of coverage when aspects of the HRQoL are not appropriately captured all come at the cost of reduced cross program comparability (Brazier et al, 2012; Brazier and Tsuchiya, 2010; Dowie, 2002; Moock and Kohlmann, 2008). Using a different GPBM from the reference case implies that different patients are assessed in terms of different health dimensions and valuation methods (Brazier et al, 2011) (see the description of GPBMs in this chapter for more details). Using a condition specific measure introduces a number of sources of non-comparability even where valuation methods are the same e.g. TTO values from general population. For example, naming the condition (as usually done in condition specific measures) might influence the valuation due to prejudices for the condition e.g. shortness of breath due to asthma might be valued differently from shortness of breath due to cancer (Brazier and Tsuchiya, 2010). Similarly, the narrower descriptive system of condition specific measures may miss dimensions that are relevant to describe health. These missing dimensions might interact with those present in the measure i.e. preference interaction creating biases in the utility values obtained (Brazier and Tsuchiya, 2010). Finally, use of condition specific measures has been seen to cause focusing effects in responders, which occur when values for a given health state are low as the responder focuses on few problems, but if he or she had been given fuller information on other aspects of health his or her judgments would be improved (Brazier et al, 2017).

In response to these problems it has been proposed to add bolt-ons to the descriptive system of the GPBMs of interest (EuroQol, 2012). Bolt-ons are dimensions that can be added to a GPBM. They were initially conceived to overcome the inadequacies of the parent instrument in a specific population (Longworth et al, 2014), but can also be used to extend a measure into a broader measure of health, or a broader measure covering areas beyond health. The addition of bolt-on dimensions produces an extended descriptive system defined by the combination of the original dimensions plus the bolt-on. This results in a number of new health states equal to the number of original health states multiplied by the number of bolt-on levels. Health state values can be subsequently obtained through preference elicitation techniques such as TTO and SG. The use of bolt-ons allows the content validity of the measure to be improved, while simultaneously maintaining its core structure (Swinburn et al, 2011). Therefore bolt-ons may restore some form of comparability between assessments. The next section reviews the bolt-on studies conducted to date for the EQ-5D.

3.5 A review of bolt-on studies for the EQ-5D

Research into the development of bolt-ons has been largely exploratory to date and has only focused on the EQ-5D (Brazier et al, 2017). In order to collect and synthetize the evidence on bolt-ons developed to date and their impact, a scoping review was conducted. Its methods and results are reported below.

3.5.1 Methods

A databases free text search (Higgins and Green, 2011) was performed to identify a set of key papers (Higgins and Green, 2011) relevant for the topic investigated. This type of search was chosen as a primary search as it is the most efficient way to retrieve published literature in health related topics (Higgins and Green, 2011).

The free text search investigated Medline (Ovid), Scopus and Cochrane Library, combining two concepts, "EQ-5D" and "bolt-ons", both with spelling variations and acronyms, without date or language restrictions. The search terms for Medline (Ovid) were:

1. euroqol OR "euro qol" OR eq5d OR "eq 5d" OR eq-5d OR eq5d5l OR "eq 5d 5l" OR eq-5d-5l OR "euro adj qol" OR "eur adj qual" OR eq adj 5d.tw.

2. "bolt-on" OR "bolt on" OR "bolt ons" OR "bolt-ons" OR "add-on" OR "add-ons" OR "add adj ons"

3. 1 AND 2

The searching strategy was adapted for differences across databases.

A set of pre-defined eligibility criteria was used to screen the references. Identified records were included if they tested the impact of adding one or more dimensions to the EQ-5D descriptive system, they investigated the development of a disease specific version of the EQ-5D, they investigated the development of an extended version of the EQ-5D or they tested the impact of the EQ-5D + bolt-ons. Records were excluded if they did not mention the EQ-5D in the title or abstract and if they investigated the psychometric performance of the EQ-5D.

Subsequently, the reference list of included articles was screened to retrieve additional records. This secondary strategy was considered necessary for two main reasons. First, bolt-on studies may fail to mention the word "bolt-on" or "add-on" and simply refer to the specific dimension being bolted on e.g. development of the EQ-5D + fatigue. When this happens, even a well-designed and comprehensive database search would fail to retrieve some of the relevant literature. Second, commonly used terms such as "dimensions" or "items" are often employed as synonyms of bolt-ons. Inclusion of these terms in the database search would have substantially decreased its specificity (Edwards et al., 1998, pg. 257-258; Dolan et al., 2005), making the search unfeasible.

Potentially relevant records obtained from the reference lists of included articles were assessed based on the same criteria of the database search. Whenever an additional record was included, its reference list was screened in search of new records. This iterative process was considered completed once no additional potentially relevant record was found.

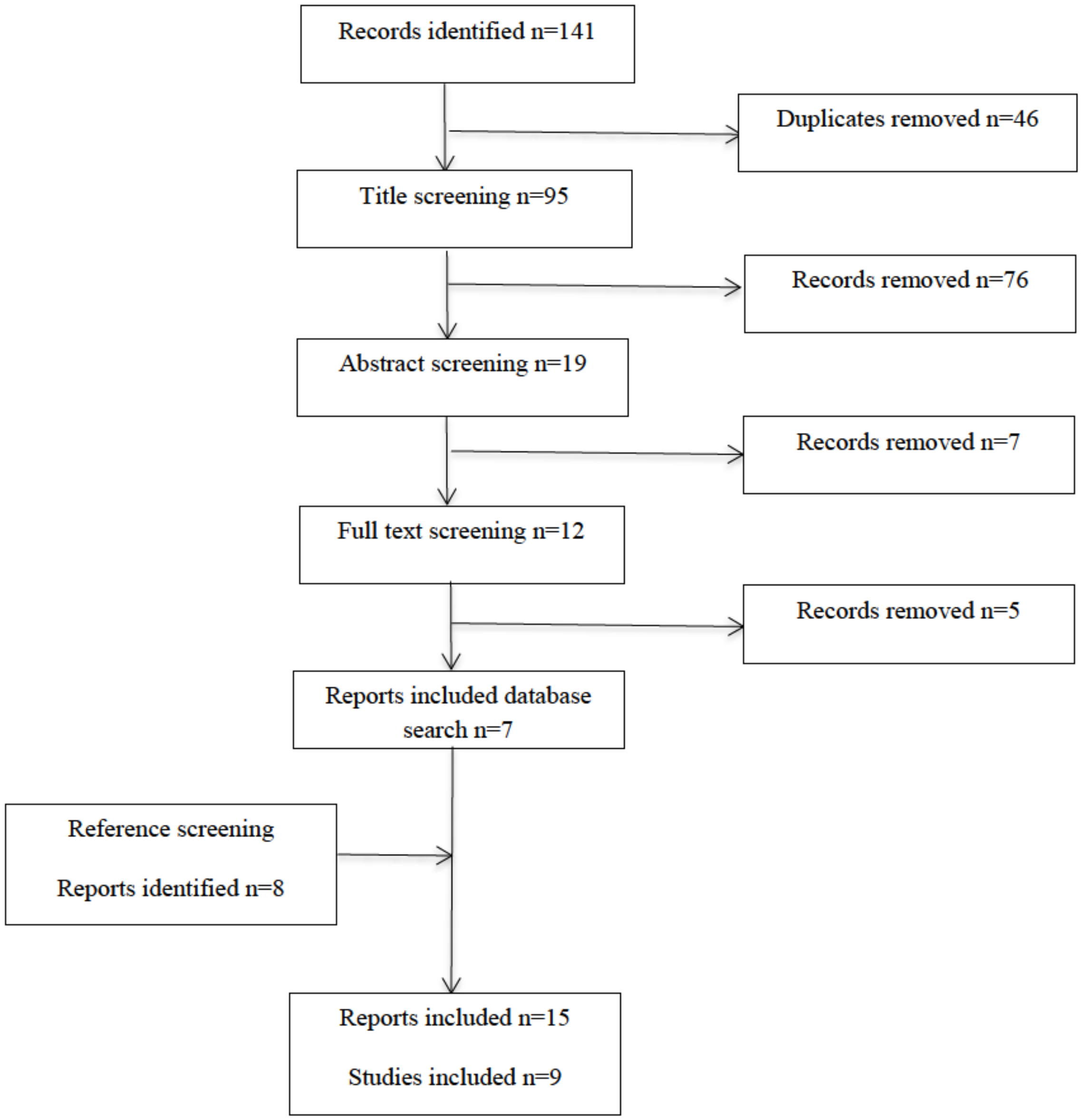
Information from the included records was extracted using a standardized extraction template. Whenever a study was reported in multiple reports, the article with the most up-to-date data was used as a main source and supplemented by additional evidence contained in the other sources. Results are summarized in a narrative form.

3.5.2 Results

The database search retrieved a total of 141 records, 46 of which were duplicates and were for this reason removed. Seventy-six records were removed after title screening, 7 after abstract screening and 5 after full text screening. The remaining 7 records were included in the review. Reference screening identified 8 additional records. Hence, the final selection included a total of 15 reports related to nine studies. Included studies were Gudex (1991), Krabbe et al (1999), Yang et al (2014), Yang et al (2015), Swinburn et al (2013), Hogendoorn et al (2016), Luo et al (2015), Wolfs et al (2007), Arons and Krabbe (2011). A flow chart of the selection process is presented in figure 3.2.

Six of the included studies investigated the development of one or more bolt-on dimensions (Gudex, 1991; Krabbe et al, 1999; Yang et al, 2014; Yang et al, 2015; Swinburn et al, 2013; Hogendoorn et al, 2016), while three of the included studies the psychometric testing of a version of the EQ-5D + bolt-on (Luo et al, 2015; Wolfs

Figure 3.2 – Flow chart of the selection process



et al, 2007; Arons and Krabbe, 2011). Two studies used the 5L version of the EQ-5D (Hoogendoorn et al, 2016; Swinburn et al, 2013) and 7 studies the 3L version (Gudex, 1991; Krabbe et al, 1999; Yang et al, 2014; Yang et al, 2015; Luo et al, 2015; Wolfs et al, 2007; Arons and Krabbe, 2011).

Table 3.3 presents a summary of the main characteristics of the 6 studies investigating the development of a bolt-on. As it can be seen, four studies tested the impact of adding one bolt-on (Gudex, 1991; Krabbe et al 1999; Yang et al, 2014; Hoogendoorn et al, 2016), one assessed, comparatively, the impact of 2 bolt-ons and another one assessed, comparatively, the impact of 3 bolt-ons (Yang et al, 2015). Bolt-ons developed were energy, cognition, sleep, vision, hearing, tiredness, skin irritation, self-confidence and breathing.

Studies differed in the evidence used to identify bolt-on dimensions. Three studies relied on only one type of evidence, and this was expert opinion in Krabbe et al (1999), previous results from a large scale survey in Gudex (1991) and literature reviews of psychometric evidence in Hogendoorn et al (2016). The remaining three studies relied on multiple types of evidence. One of them used expert opinion, presence of the dimension in other commonly used GPBMs and prevalence of the problem in the British population (Yang et al, 2014); another study used literature reviews of psychometric evidence and expert opinion (Yang et al, 2015); and a third study used literature reviews of psychometric evidence, literature review of the aspects of health that impact the condition of interest and qualitative interviews with patients and experts (Swinburn et al, 2013).

Differences were also found in the process followed to select item/s related to the bolt-on dimensions of interest and develop dimensions. Three studies did not explain how items were identified, but explained that these were developed to reflect a similar operationalization to the one used in the parent measure (Krabbe, 1999; Yang et al, 2014; Yang et al, 2015). One study used qualitative interviews to draft two dimensions wording, and tested them using a cognitive debriefing exercise (Swinburn et al, 2013). Another study developed two items for the bolt-on dimension of interest using a process that involved a content review of six disease specific measures, regression analysis to investigate the proportion of variance explained by items included in disease specific questionnaires and a principal component analysis of the EQ-5D + disease specific measures (Hogendoorn et al, 2016). One study did not explain the process followed for developing the bolt-on dimension (Gudex, 1991).

All studies tested whether a bolt-on had an impact on the EQ-5D by means of comparing the values obtained from a standard version of the questionnaire with the values obtained from the questionnaire + bolt-on dimensions. Differences were found in the elicitation techniques used to derive health state utility values. Two studies used VAS (Gudex, 1991; Krabbe, 1999), three studies used TTO (Yang et al, 2008; Yang et al, 2014; Longworth et al, 2014; Yang et al, 2015) and one study used TTO and discrete choice experiments (Hogendoorn et al, 2016).

Most of the bolt-ons tested appeared to have an impact on health state values for the EQ-5D. More specifically, 7 of the 9 bolt-ons i.e. cognition, vision, hearing, tiredness, skin irritation, self-confidence and breathing were found to alter health state utility values in the expected direction, with differences between the questionnaire without bolt-on and the questionnaire with bolt-on being generally statistically significant. Two bolt-ons were found to have little or no impact on health state utility values i.e. sleep and cognition.

Table 3.4 presents the main characteristics of the three studies that tested the psychometric properties of the EQ-5D + bolt-ons. One study investigated the performance of the EQ-5D + vision (Luo et al, 2015), while two studies the performance of the EQ-5D + cognition (Wolfs et al, 2007; Arons and Krabbe, 2011). The study investigating the performance of the EQ-5D + vision performed a known groups assessment. The other two studies assessed the responsiveness of the measure, as well as the convergent validity with the mini mental state examination at baseline, 6 months and 12 months.

The study by Luo and colleagues (2015) found a better performance of the EQ-5D + cognition in patients affected by visual disorders. The two studies investigating the performance of the EQ-5D + cognition differed in their conclusions, as Arons and Krabbe (2011) reported an improved responsiveness and convergent validity of the EQ-5D + cognition compared to the standard EQ-5D, while Wolfs and colleagues (2007) concluded that the EQ-5D performed sufficiently well without the addition of cognition. However, this latter conclusion does not seem entirely justified by the results reported by the authors, as correlations between the EQ-5D + cognition and the mini mental state examination were higher than correlations between the standard EQ-5D and the mini mental state examination at all endpoints examined.

3.5.3 Discussion

The current scoping review found that 9 bolt-ons have been developed to date for the EQ-5D. These are cognition, energy, vision, hearing, tiredness, skin irritation and self-confidence, breathing and sleep. Seven of these bolt-ons had an impact on health state utility values, while 2 had no or limited impact. Two bolt-ons, vision and cognition, have undergone further psychometric testing in three studies (Luo et al, 2015; Wolfs et al, 2007; Arons and Krabbe, 2011). Vision was found to improve the known group validity of the EQ-5D compared to the standard questionnaire. Cognition was found to improve the convergent validity and responsiveness in one study, and to provide a marginally improved convergent validity in another study. Overall, these results seem to suggest that bolt-ons are an interesting avenue of further study.

Most of the studies included in the current scoping review identified bolt-ons based on expert opinions and literature reviews of psychometric evidence. Expert opinion are easy to use and quick to collect, but are by definition subjective. Literature reviews of psychometric evidence may be affected by numerous problems as shown in the overview of reviews presented earlier in this chapter. Among them, reviews often do not adhere to the quality standard required. Moreover, studies included in literature reviews of psychometric evidence may ignore the specific nature of these measures, performing incorrect tests, using the wrong indicators e.g. clinical indicators or drawing wrong conclusions e.g. rarely GPBM are tested on the unscored dimensions as it is recommended by Brazier and Deverill (1999). Furthermore, even when the correct tests are reported, studies included in literature reviews of psychometric evidence may report p values and confidence intervals inconsistently, as well as not stating clearly the thresholds used to draw conclusions. More importantly, literature reviews rarely report results for the unscored dimensions of GPBMs, therefore not allowing the identification of specific dimensions missing from the investigated measure.

An alternative method to identify bolt-ons that was not covered in the current scooping review has been proposed by Lin and colleagues (2013). This consists in comparing the content of the EQ-5D with the content of condition specific measures, identifying relevant condition specific measures using a review of mapping studies. A major limitation of this method for the purpose of identifying bolt-ons is that it does not clarify whether the constructs measured by conditions specific measures are independent and whether they contribute in explaining HRQoL variance. Overall, both from the studies included in the review and this additional study it emerges that the methods used to date for identifying bolt-ons might not be appropriate, and that an effort should be put in exploring more robust techniques for this purpose. Studies included in the current review employed different approaches to select between items and test the impact of the developed bolt-ons. One study selected items based on linear regressions where the ability of items to predict variations in HRQoL was assessed (Swinburn et al, 2013). Another study, which was conducted after the beginning of this research program, employed principal component analysis to assess the unidimensionality of the items developed (Hogendoorn et al, 2016). The remaining studies used information from preference elicitation techniques to determine whether a bolt-on had an impact on preference for the EQ-5D (e.g. Yang et al, 2015). These approaches have never been tested before for selecting bolt-on dimensions. It is for this reason unclear whether they are appropriate for this purpose, and if they are how they do compare.

3.6 Research gap

The scoping review of bolt-on studies has evidenced that techniques used to date might not be appropriate to identify the bolt-on dimensions missing from the descriptive system of a GPBM. Moreover, the scoping review has noticed that the techniques employed to select bolt-ons have never been tested for this purpose. The lack of clarity on whether these techniques are the only possible alternatives for identifying bolt-on dimensions, and whether they can be used to select between bolt-ons represents a major limit that undermines the development of the field.

3.7 Study aims and objectives

The main aim of this thesis is to explore methods for identifying potentially relevant bolt-on dimensions, and to examine the usefulness of two methods for selecting between bolt-ons. As it would be unfeasible to investigate these issues on all the existing GPBMs, the EQ-5D-5L has been chosen as a case study. However, methods explored in this thesis will be generalizable also to other measures.

In order to meet this aim, the following objectives will need to be addressed:

1. To examine the usefulness of employing quantitative methods such as exploratory factor analysis, principal component analysis, confirmatory factor analysis and structural equation modelling for the purpose of bolt-on identification;
2. To examine whether it is possible to use the ability of bolt-ons to predict differences in HRQoL to select between them;
3. To examine whether it is possible to use the impact of bolt-ons on preference for pairs of health states to select between them.

3.8 Conclusions

This chapter has introduced the methods commonly used for developing GPBMs and has described important decisions that measures' developers need to face when selecting dimensions, items and levels for the descriptive system of their measure. In doing this, it has provided a pragmatic overview of the choices that lead different GPBMs, all aiming at measuring HRQoL, to include different dimensions and items. It has also introduced the importance of testing for the face validity and relevance of the dimensions and items, an aspect that will be further investigated in Chapter 6. It has explained that despite consistent decisions can be achieved only using the same GPBM in all assessments this might not be appropriate as the chosen measure might lack validity and responsiveness in the condition or disease area of interest. It has for this reason summarized all reviews on the performance of GPBMs, finding that despite these measures are generally valid and responsive, they all report problems of validity and responsiveness in some conditions or disease areas. There are different solutions commonly employed when the GPBM of interest lacks validity and responsiveness in a condition or disease area, and these have been explained. One of these solutions, using bolt-ons, appear preferable to others as it ensures comparability between assessments. For this reason, the chapter has reviewed through a scoping review the bolt-ons studies conducted to date, finding that bolt-ons have been identified relying on literature reviews and expert opinion, and selected using methods such as regressions and their impact on preferences. There are numerous problems with the use of expert opinion as this method is arbitrary. There are also problems with the use of literature reviews of psychometric evidence, as these are often poor in quality, use methods not appropriate to test GPBMs and rarely report information for the dimensions of the mesaures. Also, methods used for selecting bolt-ons have never been explored for this purpose, and it is therefore not clear their usefulness. This represents a major gap that undermines the development of the field. For this reason, the next chapters will explore alternative methods for identifying potentially relevant bolt-on dimensions, and the usefulness of two methods for selecting between bolt-ons.

Table 3.1 Characteristics of the included reviews

Disease area	*Report*	*Condition/ population*	*Quality*	*GPBM/s included*	*Aims and objectives*	*Number of reports (number of studies)**
Autoimmune system	Castelino (2013)	Systemic lupus erythematosus	Poor	EQ-5D SF-6D	To evaluate the development and psychometric properties of health related quality of life measures used in adults with systemic lupus erythematosus	13 (13)
	Holloway (2014)	Systemic lupus erythematosus	Poor	EQ-5D	To create a conceptual model of the humanistic and economic burden of systemic lupus erythematosus and review the patient reported outcomes used to measure the concept in SLE clinical trials	68 (68)
Cardiovascular system	Dyer (2010)	Heart disease	Good	EQ-5D SF-6D HUI 3	To synthetize the evidence on the validity and reliability of the EQ-5D in studies within the cardiovascular field; to summarize the EQ-5D based score reported in studies within the cardiovascular field; and to attempt to stratify mean utility	66 (66)

					scores according to level of disease severity	
Ear	Yang (2013)	Hearing impairment	Good	EQ-5D SF-6D HUI 3	To assess the reliability, validity and responsiveness of the EQ-5D, HUI 3 and SF-6D for measuring health related quality of life in people with hearing impairment	18 (14)
Endocrine, nutritional and metabolic diseases	Janssen (2011)	Type 2 diabetes	Good	EQ-5D SF-6D 15 D	To summarize the evidence on the validity, reliability and responsiveness of the EQ-5D in studies of diabetes type 2	59 (59)
	Speight (2009)	Type 2 diabetes	Poor	EQ-5D	To clarify the measurement of QoL in terms of conceptualization, terminology and psychometric properties, to review the instruments that have been most frequently used to assess QoL in diabetes and make recommendations in how to select measures appropriately	19 (19)
Eye	Tosh (2012)	Visual	Good	EQ-5D	To assess the appropriateness of	31 (31)

		impairment		SF-6D HUI 3	the EQ-5D, HUI 3 and SF-6D in patients with visual disorders due to the different ways particular conditions affect HRQoL	
Genitourinary system	Davis and Wailoo (2013)	Urinary Incontinence	Good	EQ-5D SF-6D 15 D AQoL 8D	To assess the appropriateness of the EQ-5D in people with urinary incontinence	17 (17)
	Wu (2013)	HIV	Good	EQ-5D	To examine the responsiveness of two health related quality of life measures used in clinical trials involving HIV infected adults	17 (17)
Gynaecological problems	Sanghera (2013)	Menorrhagia	Poor	EQ-5D	To review which economic measures have been used or assessed in menorrhagia and present criteria for deciding which measure is the most appropriate	56 (56)

Haematological problems	Szende(2003)	Haemophilia	Good	EQ-5D HUI 3	To review and evaluate the performance of health related quality of life and other health status measures used in studies of haemophilia in adult patients and provide recommendations for future research	19 (19)
Musculoskeletal system	Bansback (2008)	Rheumatoid arthritis	Poor	EQ-5D SF-6D HUI 3	To review the clinical measures used in rheumatoid arthritis economic evaluations with respect to their relevance and sensitivity to changes in survival, health related quality of life and costs	22(22)
	DeVine (2011)	Chronic low back pain	Poor	EQ-5D	To determine the correlation of patient reported pain with physical function and health related quality of life after spine surgery and the responsiveness of pain, physical function and health related quality of life measures after spine surgery (for chronic low back pain)	5 (5)

	Hill (2010)	Spinal cord injury	Poor	SF-6D	To critically review quality of life instruments used in spinal cord injury	14 (14)
	Whitehurst (2012)	Spinal cord injury	Good	SF-6D	To review the use of generic-preference based instruments of health related quality of life within the context of spinal cord injury	22 (22)
Mental health	Brazier (2014)	Bipolar disorder	Good	EQ-5D	To examine the validity and responsiveness of two generic preference based measures of health (the EQ-5D and SF-6D) and two generic non preference based measures (the SF 36 and SF 12) in populations with bipolar disorder	22 (22)
	Papaioannou (2013)	Personality disorder	Good	EQ-5D	To assess the construct validity and responsiveness of four generic health status measures in personality disorder	10 (10)
	Papaioannou (2011)	Schizophrenia	Good	EQ-5D	To assess the construct validity and responsiveness of four generic health status measures in	33 (33)

				SF-6D	schizophrenia	
	Peasgood (2012)	Depression/ Anxiety	Good	EQ-5D SF-6D HUI 3	To assess the construct validity and responsiveness of EQ-5D and SF-6D measures in depression and anxiety	26 (26)
	Hounsome (2011)	Dementia	Poor	EQ-5D HUI 3	To review evidence relating the application of EQ-5D in dementia research and issues concerning its use	21 (18)
Neoplasm	Longworth (2014)	Cancer	Good	EQ-5D SF-6D HUI 3	To assess the reliability, validity and responsiveness of the EQ-5D, HUI 3 and SF-6D for measuring health related quality of life in cancer	98 (98)
	Pickard (2008)	Cancer	Poor	EQ-5D	To summarize evidence on the validity and reliability of EQ-5D in cancer	34 (34)
Nervous system	Kuspinar and Mayo (2014)	Multiple Sclerosis	Excelle nt	EQ-5D SF-6D HUI 3	To summarize the evidence from published literature on the psychometric properties of a generic utility measures in	15 (15)

				AQoL 8D	multiple sclerosis	
Nose	Linder (2003)	Acute sinusitis	Excelle nt	EQ-5D	To identify and compare the performance of HRQoL instruments or symptom scores for adults with acute sinusitis.	29 (29)
Others	Ching (2003)	Aesthetic surgery	Poor	EQ-5D	To critically review the present literature to identify the appropriate instruments to assess outcomes in aesthetic surgery.	43 (not clear)
	Derrett (2009)	Injuries	Poor	EQ-5D	To describe EQ-5D administration, summarize its reliability and validity and report its outcomes in injuries	44 (41)
	Haywood (2005)	Older patients	Poor	EQ-5D AQoL 8D	To review the evidence relating to the measurement properties of multi-item generic patient or self assessed measures of health in older people	122 (122)
Respiratory system	Petrillo (2011)	Asthma/COPD	Poor	EQ-5D	To present and discuss the empirical evidence on the validity of generic multi attribute utility instruments within the	22 (22)

					COPD population	
	Pickard (2007)	Asthma/COPD	Good	EQ-5D SF-6D	To synthetize literature on the validity and reliability of EQ-5D use in studies of asthma and COPD, and estimate EQ-5D utility scores associated with different stages of the disease	18 (18)
Skin and subcutaneous tissues	Yang (2014)	Skin condition	Good	EQ-5D	To assess the reliability, validity and responsiveness of the EQ-5D, HUI 3 and SF-6D for measuring health related quality of life in skin conditions	16 (16)

* Note: Not all the studies included in the reviews were relevant to the research question investigated in this overview. However, this overview draws on more than hundred and fifty studies included in the thirty reviews.

Table 3.2 Main EQ-5D, SF-6D and HUI 3 results

Disease Area/Population		Authors	Condition examined or population examined	Known groups			Convergent validity			Responsiveness		
				EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III
	Autoimmune system	Castelino (2013)	Systemic Lupus Erythematosus			N/R	✓	±	N/R	✓	✓	N/R
		Holloway (2014)	Systemic Lupus Erythematosus	✓	N/R	N/R	✓	N/R	N/R		N/R	N/R
	Cardiovascular system	Dyer (2010)	Cardiovascular diseases	✓ ✓✓✓ ✓ ✓✓✓ ±✗ ✗✗			✓ ✓✓✓ ✗ ✗	✓✗	✓✗	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓±±± ✗ ✗✗✗ ✗ ✗✗✗ ✗///		
	Ear	Yang (2013)	Hearing impairment	✗		✓ ✓✓✓ ✗	✓/	✓	✓ ✓//	✓✗ ✗✗		✓ ✓✓✓ ✓
	Endocrine, nutritional and metabolic system	Janssen (2011)	Type 2 Diabetes	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✗ ✗	✓✗	N/R	✓ ✓✓✓ ✓ ✓± ✗ ✗			✓ ✓✓✓ ✓✗/		N/R
		Speight	Diabetes		N/R	N/R		N/R	N/R	✗	N/R	N/R

	Authors	Condition examined or population examined	Known groups			Convergent validity			Responsiveness		
			EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III
	(2009)										
Eye	Tosh (2012)	Visual disorders	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓±±±± ±± ✗ ✗	✓	✓ ✓	✓ ✓✓✓ ± ✗ ✗✗ ✗	✓	✓	✓± ✗		✓
Genito-urinary system	Davis and Wailoo (2013)	Urinary Incontinence	✓ ✓✓✓ ±	✓	N/R	✓ ✓✓✓ ✓ ✓✓//		N/R	✓ ✓✓✓ ✓ ✓✓±		N/R
	Wu (2013)	HIV	✓	N/R	N/R		N/R	N/R	✓ ✓✗ ✗ ✗	N/R	N/R
Gynaecological problems	Sanghera (2013)	Menorrhagia	✓	N/R	N/R	✗ ✗	N/R	N/R	✗	N/R	N/R
Haematological problems	Szende(2003)	Haemophilia	✓ ✓	N/R	✓	✓	N/R			N/R	

	Authors	Condition examined or population examined	Known groups			Convergent validity			Responsiveness		
			EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III
Musculoskeletal system	Bansback (2008)	Rheumatoid arthritis				✓	✓	✓			
	DeVine (2011)	Chronic Low Back Pain		N/R	N/R	/	N/R	N/R	/	N/R	N/R
	Hill (2010)	Spinal Cord Injury	N/R		N/R	N/R		N/R	N/R	✓	N/R
	Whitehurst (2012)	Spinal Cord Injury		✓ ✓✓✗ ✗ ✗✗						✓	
Mental Health	Brazier (2014)	Bipolar disorder	✓ ✓✗	N/R	N/R	✓ ✓✓✗ ✗	N/R	N/R		N/R	N/R
	Papaioannou (2013)	Personality Disorder	✓ ✓✗	N/R	N/R	±±	N/R	N/R	✓ ✓/	N/R	N/R
	Papaioannou (2011)	Schizophrenia	✓		N/R	✓±±± ✗ ✗ ✗✗	✗	N/R	✓✗/		N/R
	Peasgood (2012)	Depression and Anxiety	✓ ✓✓✓ ✓ ✓✓✓ ± ✗	✓ ✓✓ ✓ ✓±	✓	✓ ✓✓✓ ✓ ✓	✓±	✓	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓± ✗//	✓ ✓✓	
	Hounsome (2011)	Dementia		N/R		✓ ✓✓✓ ±± ✗ ✗	N/R	✓		N/R	
Neoplasm	Longworth (2014)	Cancer (various)	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓	✗	✓ ✓✓✓ ✓ ✓✓✓ ± ✗ ✗	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✗ ✗ ✗/	✓	✓ ✓✓✓ ✗ ✗	✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓ ✓ ✓✓✓		✓ ✓✓✓ ✓ ✓✓✓ ±±

	Authors	Condition examined or population examined	Known groups			Convergent validity			Responsiveness		
			EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III
			✓✓✓✓ ✓✗✗✗ ✗✗✗						✓✓✓✓ ✓✓✓✓ ✓±±±± ±±✗✗ ✗✗✗//		
	Pickard (2008)	Cancer (various)	✓✓✓✓ ✓✓✓✓	N/R	N/R	✓	N/R	N/R	✓✓	N/R	N/R
Nervous system	Kuspinar and Mayo (2014)	Multiple Sclerosis	±±✗✗	±	✓	✓✓✓± ±✗	✓✓	✓			
Nose	Linder (2003)	Rhino sinusitis		N/R	N/R		N/R	N/R	✗	N/R	N/R
Others	Ching (2003)	Aestetic surgery		N/R	N/R		N/R	N/R	±	N/R	N/R
	Derrett (2009)	Injuries	✓✓✓✓	N/R	N/R	✓✓✓✓ ✓±✗	N/R	N/R	✗	N/R	N/R
	Haywood (2005)	Older population		N/R	N/R	/	N/R	N/R	✓	N/R	N/R
Respiratory system	Petrillo (2011)	COPD	✓±	N/R	N/R		N/R	N/R	✓✓✓	N/R	N/R
	Pickard (2007)	COPD and Asthma	✓✓✓✓ ✓✓✓✓ ✓✓±		N/R	✓✓✓✓ ✓±±✗	✓	N/R	✓✗✗✗		N/R

	Authors	Condition examined or population examined	Known groups			Convergent validity			Responsiveness		
			EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III	EQ-5D	SF-6D	HUI III
Skin and subcutaneous tissues	Yang (2014)	Psoriasis, acne, hidradenitis suppurativa, hand eczema, venous leg ulcers	✓✓✓✓ ✓✓✓✓ ✓	N/R	N/R	✓✓✓✓ ✓✓✗	N/R	N/R	✓✓✓✓ ✓✓✓✓ ±±✗	N/R	N/R

Legend: ✓results in support of validity or responsiveness; ✗results against validity or responsiveness; ± mixed results (some tests in support and some against); / inconclusive results (e.g. data too sparse to assess correlations;) N/R measure not reported in the review Note: Every symbol corresponds to one study. Studies reporting on more than one PBI generate more than one symbol.

Table 3.3 Summary of the main characteristics of the studies investigating the development of bolt-ons

Study ID	*Main report*	*Objective*	*Methods used for identifying bolt-on dimensions*	*Bolt-on tested*	*Sample size*	*Methods used for testing bolt-on dimensions*	*Impact of bolt-on*
1	Gudex (1991)	To investigate the impact of including an energy/ tiredness dimension to the 3L version of the EQ-5D	Previous large scale survey	Energy	340	Comparison of health status based on VAS for EQ-5D with and EQ-5D without bolt-on	No impact
2	Krabbe et al (1999)	To investigate the impact of adding a cognitive dimension to the 3L version of the EQ-5D	Expert opinion of a Dutch group of experts	Cognition	87	Comparison of health state values for paired health states based on VAS with and without bolt-on using T tests	Impact

3	Yang et al (2014)	To investigate the impact of adding a sleep dimension to the 3L version of the EQ-5D	Expert opinion, presence of the dimension in other commonly used GPBMs, importance of the aspect of health in relation to a number of conditions e.g. urinary impairments and prevalence of the problem in British population	Sleep	160	Comparison of health state values for paired health states based on TTO with and without bolt-on using T tests	No impact
4	Yang et al (2015)	To test comparatively the impact of three bolt-on dimensions on the 3L version of the EQ-5D	Literature review highlighting poor psychometric characteristics of the EQ-5D in vision and hearing. Expert opinion of NICE and its stakeholders for tiredness	Vision, hearing and tiredness	300	Comparison of health state values from paired health states based on TTO with and without bolt-ons using T test	Impact

5	Swinburn et al (2013)	To test and value a disease specific version of the 5L version of the EQ-5D	Literature review highlighting inconclusive psychometric evidence for the EQ-5D in dermatological conditions, a literature review to understand the aspects of health that impacted the condition of interest and qualitative interviews to assess whether expert in the field and patients agreed on the importance of these aspects	Skin irritation and self confidence	300	Comparison of health state values for paired health states based on TTO with and without bolt-on	Impact
6	Hogendoorn et al (2016)	To test the impact of two respiratory bolt-ons on the 5L version of the EQ-5D	Literature review highlighting poor responsiveness and face validity of the EQ-5D in patients affected by COPD	Breathing	430	Comparisons of health state values for paired health states based on TTO and DCE with and without bolt-ons using T tests and Chi square tests	Impact

Table 3.4 Summary of the included studies testing the psychometric characteristics of developed bolt-ons

Study ID	*Main report*	*Objective*	*Methods*	*Sample size*	*Results*
1	Luo et al (2015)	To assess whether the 3L version of the EQ-5D + vision discriminated better than the standard 3L version of the EQ-5D in individuals with visual disorders	Individuals were recruited in a specialist outpatient clinic in Singapore. Participants were clustered into 5 groups based on visual acuity scores. They were administered the standard 3L EQ-5D and subsequently the vision bolt-on. Known groups were paired in order to compare no versus mild, mild versus moderate, moderate versus severe and severe versus blind groups. Discriminatory power was assessed using squared T statistic. A higher T statistic shows a higher likelihood for the measure to show statistical significance when comparing groups	500	The study found higher T statistic scores for the vision bolt-on compared to the remaining dimensions of the EQ-5D in 14 of the 16 comparisons performed. Authors concluded that the vision bolt-on improves the discriminatory power of the EQ-5D in populations affected by visual disorders

2	Arons and Krabbe (2011)	To compare the performance of the 3L version of the EQ-5D with the 3L version of the EQ-5D + cognition in elderly patients	Individuals were recruited as part of a randomized controlled trial in the Netherlands. The 3L version of the EQ-5D and 3L version of the EQ-5D + cognition were collected for patients and their proxies at baseline, 6 months and 12 months. Utilities for both the EQ-5D and the EQ-5D + bolt-ons were calculated. ANOVAs were used to assess difference between utilities between baseline, 6 months and 12 months. The content validity of the cognition dimension was assessed regressing the EQ-5D on the Health VAS and subsequently adding cognition	175	Utilities of the EQ-5D + cognition differed significantly between baseline and 6 months and between baseline and 12 months. Differences for the standard version of the EQ-5D were not significant. The cognition dimension reported statistically significant results and its coefficient were larger than for most other EQ-5D dimensions. Authors concluded that the addition of a cognition dimension improves the content validity and responsiveness of the EQ-5D

3	Wolfs et al (2007)	To compare the construct validity and responsiveness of the 3 L version of the EQ-5D and the 3L version + cognition in elderly patients	Individuals were recruited as part of a randomized controlled trial in the Netherlands. The 3 L version of the EQ-5D was initially collected from the individuals' proxies. Subsequently, the mini mental state examination and the cognition dimension were administered. Spearman's rank correlations were used to analyse the association between the EQ-5D and the mini mental state examination, and the EQ-5D + cognition and the mini mental state examination at baseline, 6 and 12 months with	234	Authors found that correlations between the EQ-5D overall score, the individual dimensions and the mini mental health state were generally moderate. Also the correlation between the cognition dimension and the mini mental health state was moderate. Although, this was higher than the correlation of both the overall idenx and the individual dimensions at baseline, 6 months and 12 months, authors concluded that the EQ-5D performs sufficiently well without the addition of cognition

Chapter 4

Using Principal Component Analysis, Exploratory Factor Analysis, Confirmatory Factor Analysis and Structural Equation Modeling to identify bolt-on dimensions

4.1 Introduction

Chapter 3 has reviewed the existing literature on the psychometric performance of the 5 mostly used GPBMs, across all disease areas and conditions. The chapter found that GPBM are mostly valid, but that all of them might report problems of validity and responsiveness in some conditions or disease areas. It also found that this type of evidence is often limited in terms of its breath and depth, which makes it difficult to prove one way or another that a measure is not valid, and that evidence included in literature reviews of psychometric evidence is affected by some methodological problems. The chapter has also reviewed the bolt-on literature, finding that bolt-ons have been identified using expert opinion and literature reviews of psychometric evidence.

Given that expert opinion is arbitrary by definition, and literature reviews of psychometric evidence are affected by some methodological problems previously mentioned, the current chapter investigates the use of a secondary data analysis as an alternative for identifying bolt-ons. This might represent a better technique as it allows to identify the specific dimensions missing from the investigated GPBM relying upon empirical data.

As previously mentioned in this thesis (Chapter 3), GPBMs differ in terms of their content (Brazier et al, 2012). While the EQ-5D focuses on physical health (4 of its 5 items cover mobility, usual activities, self-care and pain/discomfort), other measures such as the AQoL 8D report a prevalence of psychosocial constructs (25 of its 35 items regard constructs, among the others, of anger, self-esteem, satisfaction and intimacy). Five GPBMs are commonly used in economic evaluations, and these are the EQ-5D, SF-6D, HUI 3, AQoL 8D and 15D. These have been shown by

Richardson and colleagues (2015) to cover most of the health domains of interest for the HRQoL of individuals and their items and the dimensions they relate to might therefore represent an important pool of candidate bolt-ons. For this reason, these measures were examined in the current secondary data analysis. In addition, four validated and frequently employed subjective wellbeing measures (SWBMs) were used in this study, as these might cover additional potentially relevant constructs for bolt-ons identification (see discussion on subjective wellbeing in Chapter 2). These are the Personal well-being index (PWI), the Satisfaction with Life Scale (SWLS), the Office for National Statistics (ONS) measure and the ICEpop CAPability measure (ICECAP).

Various statistical techniques can be used to identify which items, and latent constructs, are potentially missing from a specific GPBM such as the selected EQ-5D where a large pool of items from other measures is available. These are principal component analysis (PCA), exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and structural equation modeling (SEM). PCA and EFA are data reduction methods that investigate the correlation matrix of a set of observed variables in order to reduce them into a smaller set of constructs, called components in PCA and factors in EFA (Field, 2009); CFA is a confirmatory approach that verifies the appropriateness of a measurement model (relationships between factors and indicators) derived from theory and/or preliminary empirical research (Field, 2009); SEM is a confirmatory method that expands on a CFA by determining the suitability of a structural model (regressions between factors) derived from theory and/ or preliminary empirical research (Kline, 2016). PCA and EFA may be suitable for exploring the underlying dimensional latent structure to which the item pool of the GPBMs relates and which of the constructs are covered by the EQ-5D descriptive system. CFA could then be used to confirm this anticipated dimensional structure, while SEM can help to further investigate the relationships between latent constructs. Using SEM might help identifying whether latent factors are related and may help reducing double counting. Once latent constructs i.e. factors and components and items are identified, these can be subsequently developed or adapted into actual bolt-on dimensions, as done for factors in Chapter 6.

4.2 Methods

4.2.1 Data

The study reported here used the Multi Instrument Comparison (MIC) database, the largest dataset on health and well-being measures available worldwide (Richardson et al, 2012 a). A detailed description of data collection methods can be obtained from elsewhere (Richardson et al, 2012 a; Richardson et al, 2012 b). An online survey was carried out in six countries (Australia, Canada, Germany, Norway, United Kingdom and United States) imposing quotas to obtain similar socio-demographic characteristics across countries. Responders were members of the general public who had previously agreed to participate to online surveys. Edit procedures were applied to improve data quality. Respondents were excluded if: i) they completed the survey in less than 20 minutes; ii) they stated not to have a health problem but reported a self assessed health status below 65 on the Health VAS scale; iii) large differences were found between duplicated questions; iv) more than 2 response level difference were found in pain questions. The final sample comprised 8022 individuals.

4.2.2 Measures and items

The five chosen GPBMs for this secondary data analysis are the EQ-5D-5L, the SF-6D, the HUI3, the AQoL 8D and the 15D. The four selected SWBMs are the PWI, the SWLS, the 4 item ONS measure and the ICECAP. This resulted in a pool of 92 items, 69 of which taken from GPBMs and 23 of which taken from SWBMs. Wording for the items used is available in Appendices Chapter 4 – Table I.

All items are ordinal categorical, with the number of categories varying between 4 and 11. In EQ-5D-5L, SF-6D, HUI 3, AQoL 8D, 15 D and ICECAP lowest scores represent the best possible health status (e.g. level 1 mobility of the EQ-5D-5L is perfect mobility). In PWI, SWLS and ONS lowest scores represent the worst possible health status (e.g. level 0 satisfaction with life of the ONS represent "not at all satisfied").

4.2.3 Content analysis

To provide a theoretical guide for interpreting the results, a content analysis of the items was performed following the Wilson and Cleary conceptual model (W&C) (Wilson and Cleary, 1995). The W&C model combines biological and psychological aspects of HRQoL, defining five main areas including physiological factors, symptom status, functioning status, general health and overall quality of life. A causal relationship is assumed in the model from physiological factors and symptoms to functioning to general health and overall quality of life.

GPBMs and SWBMs items were firstly clustered into homogeneous themes based on their content. Subsequently, themes were assigned to one of the five categories of the model. To reflect domains commonly measured by GPBMs and as suggested by the W&C model, the symptoms category was further classified into physical or psychological symptoms, and the functioning category into physical functioning, psychological functioning, social functioning and role functioning. The use of the W&C model was considered particularly useful for informing the SEM model and a discussion about whether a health state classification should contain items across categories.

4.2.4 Multivariate statistical analyses

The MIC sample (n=8022) was split into two random halves with 4011 observations each. One random half was used for PCA and EFA, and for specifying CFA and SEM models; the other random half for CFA and SEM cross-validation. Before extending the CFA into a full SEM, the robustness of the model was tested using 10 random resampling of the dataset.

Principal component analysis and Exploratory factor analysis

PCA and EFA are commonly used to investigate the latent structure to which a set of observable variables relate by explaining the maximum amount of variance in those variables with the smallest number of components or factors possible (Field, 2009). These techniques rely on a common set of steps.

Firstly, components or factors are extracted using a computer program that examines the matrix of correlations of the observable variables and assumes that high correlations between variables imply the existence of an underlining construct (component or factor) to which they relate. In this way, all observable variables are assumed to load onto all factors/ components. PCA relies on principal components as a method of extraction. Extraction methods commonly used for EFA are robust weighted least squares, unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring.

Subsequently, the appropriate number of components or factors needs to be selected. Selection of components or factors is usually based on their eigenvalues, which represent the relative share of total variance accounted for by that factor or component. Multiple techniques can be used for this purpose (Plucker, 2003). A common method is retaining all those factors or components whose eigenvalues is greater than 1, and this is called the Kaiser rule (Kaiser, 1960). Alternatively, the magnitude of eigenvalues can be graphically plotted i.e. scree test against their ordinal number (whether they are the first eigenvalue, the second eigenvalue etc.), and only those factors that contribute to a sharp decrease in the magnitude of successive eigenvalues are retained i.e. retaining factors and components that account for a large and distinct amount of variance (Thomson and Daniel, 1996). A third option is to retain all those factors that account for a specific amount of total variance e.g. 70%. Finally, more robust approaches can be used, such as retaining factors or components using statistical simulation techniques e.g. parallel analysis (D’Agostino and Russell, 2005).

After components or factors have been selected, rotation techniques are often used to enhance the interpretability of the component or factor loadings. Rotations are a change in the coordinate of the component or factor solution that makes the pattern of loadings more pronounced and therefore clearer. Component or factor loadings are the correlation coefficients between the items and the identified component or factor. The square of components or factor loadings represents the amount of variance in the item explained by the factor or component. Two main classes of rotation techniques exist, and these are orthogonal and oblique rotations. Orthogonal rotations assume that no relation exists between the identified components or factors, while oblique rotations assume that factors or components are correlated.

Rotations result in a set of factors or components loadings that are of easy interpretability (Pedhazur and Schmelkin, 1991). Thresholds can then be used to interpret loadings.

Despite the steps and decisions required to perform PCA and EFA are the same, there has been a long lasting debate on whether the two techniques should be considered as part of the same family or completely distinct. The reason for this debate originates by the fact that while PCA assumes that all observable variables variance i.e. items variance can be explained by the components, EFA partitions the variance of the observable variables into shared and unique variance, and uses only the shared variance to extract factors. This difference has led some authors to argue in favour of the first technique (Bentler and Kano, 1990; Gorush, 1990; Mulaik, 1990), and others in favour of the second (Steiger, 1979; Velicer and Jackson, 1990 a; Velicer and Jackson, 1990 b). The debate is not yet resolved, as many books draw a distinction between the two techniques (e.g. Pedhazur and Schmelkin, 1991), but many statistical packages e.g. SPSS continue to present PCA as their default EFA option (Costelo and Osborne, 2005).

While the debate is important for methodologists and theorists, it has little relevance for empirical purposes, as differences between PCA and EFA have been seen not to alter substantive conclusions (Velicer et al, 1982; Velicer, 1977; Arrindell and Van der Ende, 1985). Nevertheless, this study used both techniques, and compared their results using various extractions, selection of factors and components and rotation methods. PCA appeared to fit the data best. As this is the norm in the literature, the default option in popular statistical software packages (Costelo and Osbrone, 2005) and is specifically geared for categorical data in SPSS (categorical principal component analysis), it was taken as the reference model. Nevertheless, results for one of the EFA models tested will be presented and discussed in comparison to the PCA model retained.

Data suitability was tested with the Kaiser-Myer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1960; Kaiser, 1974) and the Bartlett test of sphericity (Bartlett, 1950). KMO was considered appropriate if > 0.5 and Bartlett test if $p < 0.05$ (Hair et al. 1995; Tabachnick and Fidell, 2007). Examination of the data indicated that items had non-normal distributions, which is common for categorical data. However, as non normally distributed items are often treated as normally distributed in factor analysis studies, normal PCA and EFA were tested alongside techniques specifically geared to discrete ordinal values e.g. categorical PCA and categorical EFA. Analyses were conducted using IBM Corp. Released 2013. IBM SPSS Statistics for Mac, Version 22.0. Armonk, NY: IBM Corp and Mplus version 7© (Muthen and Muthen, 1998).

For EFA, different extraction methods were compared, among which maximum likelihood and principal axis factoring, the two mostly used in the literature (Costello and Osborne, 2005). PCA used principal component extraction. For both PCA and EFA, Kaiser rule (Kaiser, 1960), scree test (Thomson and Daniel, 1996) and cumulative amount of variance rule (e.g. 70%) (D'Agostino and Russell, 2005) were compared to select factors and components. However, the final selection was informed by a parallel analysis (Horn, 1965), a Monte Carlo simulation of the eigenvalues obtained by randomly generated sets of data (n=1000) of the same size (number of variables and observations), using both raw data and permutations (to account for expected non normality of distribution). Parallel analysis was chosen as it is considered a superior alternative to other techniques for selecting factors and components (Ledesma and Valero-Mora, 1999). To aid interpretability, different rotation techniques were compared, among which Promax and Geomin oblique rotations and Varimax orthogonal rotation. It was decided to use oblique rotations as these assume that factors or components are correlated, and previous studies have shown that this is the case between health constructs (Hays and Morales, 2001; Hays et al, 1998; Hays et al, 1993; Farivar et al, 2007; Konerding et al, 2009). Component and factors loadings were interpreted using cut-offs that are robust in presence of non-normal distributions, which are: loadings ≥ 0.45 are relevant; loadings ≥ 0.55 are good; and loadings ≥ 0.63 are very good (Comrey and Lee, 1992).

Confirmatory factor analysis

CFA differs from PCA and EFA in that it does not allow for all items to freely load on all factors, but it requires the investigator to impose a measurement model to the data. Imposing a measurement model implies testing and validating a set of hypothesis on the number of relevant factors, the correlation between factors and the associations between items and factors. This removes some of the arbitrariness of PCA and EFA (e.g. component rotation technique), allowing the significance of items loadings and the appropriateness of model constraint and model additions to be tested.

PCA results were used to inform on the most appropriate factor model to fit the data for CFA (after comparing them with EFA), which was tested on one of the two random halves of the MIC sample, using Mplus version 7© (Muthen and Muthen, 1998). Robust weighted least square with means and variance adjustments (WLSMV) estimator was used as suggested for ordinal categorical data (Muthen and Muthen, 1998; Beauducel and Herzberg, 2006). Model appropriateness was assessed using two practical goodness of fit indexes, the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). RMSEA was considered acceptable when ≤ 0.08 and good when ≤ 0.05 and CFI acceptable when ≥ 0.90 and good when ≥ 0.95 (Fabrigar et al, 1999; Bentler and Dudgeon, 1996; Browne and Cudeck, 1993; Steiger, 1990; Hu and Bentler, 1999). In addition, the magnitude of factor loadings, residual correlations (RC) and modification indexes (MI) were evaluated in comparison to other items and in an integrative manner. The model was revised to improve model fit by omitting poorly associated items if these were theoretically inconsistent with the factor structure identified and by specifying residual correlations (cross-loadings and local correlations). The final model was cross-validated using the second random half of the dataset. Subsequently, model robustness was tested using 10 random resampling of 50% of the observations, and one random resampling of 10%, 20%, 30% and 40% of the observations. Practical goodness of fit indices were used to assess model appropriateness.

Structural equation modelling

SEM allows regression analysis between latent factors. Unlike standard regression analysis SEM allows one dependent variable in one model equation to become an independent variable in other parts of the SEM system. SEM is also a confirmatory technique that allows different structural path model additions to be tested.

A structural model was developed using PCA, CFA and content analysis results. Factor correlations were assumed to represent causal relationships if the factors were classified under different parts of the W&C model, and simple correlations if the factors were classified within the same part of the model. Following methodologies similar to those described for CFA, the model was tested on one of the two random half of the dataset using Mplus version 7© (Muthen and Muthen, 1998), estimated using WLSMV and judged using RMSEA and CFI fit indexes, applying the same thresholds as for CFA. Re-specifications of the model were considered in the presence of low or non-significant regression coefficients between latent variables and whenever fit indices showed possible improvements that were consistent with theory. The appropriateness of the model was cross validated using the second half of the sample. Then, alternative models were developed to examine if other relationships between latent factors were plausible based on the judgments of the PhD student and suggestions of the research team. Their goodness of fit indexes and coefficients were compared to those of the final model to understand whether alternative models could fit the data equally well.

Identification and selection of bolt-ons

The aim of the PCA, EFA, CFA and SEM was to identify potential bolt-ons for the EQ-5D. As noted, the methods have different aims and generate different results, requiring different approaches for identifying bolt-ons. Potential strategies for identifying items or components/factors are presented below.

In EFA and categorical PCA components were considered as candidate bolt-ons if none of the EQ-5D items loaded on them, as this suggests no dimension of the EQ-5D covers the constructs identified. Items were considered as candidate bolt-ons if: i) they loaded with a loading ≥0.45 on components that did not contain any of the EQ-

5D dimensions, as this shows the items are not related to the EQ-5D descriptive system; ii) they reported loadings ≤0.45 on all components, as this suggests the items are poorly associated to the component structure identified.

In CFA, factors were considered as candidate bolt-ons if none of the EQ-5D items loaded on them. Items were considered as candidate bolt-ons if: i) their main loading was on factors not covered by any of the EQ-5D dimensions; ii) they were not related to the factor structure identified.

In SEM two different approaches were compared. The first one selected only those bolt-ons that were related to symptoms or functioning and that were not already covered by any of the EQ-5D dimensions. Symptoms and functioning were chosen as these aspects of health are covered by the EQ-5D descriptive system. The second one selected factors, and items loading on factors, that were only dependent variables in the causal pathways of the SEM model, if these were unrelated to the EQ-5D descriptive system. This is because dependent variables already capture the impact of independent variables.

4.3 Results

4.3.1 Content Analysis

Content analysis results are presented in Table 4.1. Physical functioning was the category with most items, 26 out of 27 of which were taken from GPBMs. By contrast, only one item was related to the "general health" category, and this was from a subjective wellbeing measure. One category (physical symptoms) covered only one theme (pain), while other categories (e.g. psychological symptoms) covered numerous themes (anxiety/depression, coping, isolation etc.). Some items were related to more than one category. For example, the AQoL 8D item "frequency of pain interfering with usual activities" was assessed as being related to both physical symptoms and physical functioning. It was placed under the pain category but it was expected to report cross-loadings.

4.3.2 Multivariate statistical analyses

Exploratory factor analysis and Principal component analysis

Data appeared suitable to perform factor analysis (KMO=0.986; Bartlett test = 0.00). Parallel analysis using both raw data and permutations supported a 9 component and a 9 factor models (Table 4.2 and Table 4.3). These explained 61.81% of variance in PCA and 64.65% of variance in EFA and appeared easily interpretable and generally consistent with the content analysis, and were for this reason retained. Components/ Factors emerging from the analysis were, in order of eigenvalue magnitude (both in PCA and EFA), physical functioning, psychological symptoms, satisfaction/contentment, pain, relationships, speech/cognition, hearing, energy/ sleep and vision.

The EQ-5D descriptive system covered three of the nine components/factors, with anxiety/ depression loading on psychological symptoms, mobility, self-care and usual activities on physical functioning and pain on pain. The remaining components/factors, namely satisfaction, speech, relationships, hearing, vision and energy/sleep were identified as candidate bolt-ons as none of the EQ-5D dimensions loaded on them.

In PCA, forty-one items loaded on the six components not covered by any of the EQ-5D dimensions with loadings ≥0.45. Of them, seventeen mainly loaded on satisfaction, six on energy/sleep, six on relationships, six on speech/cognition, three on hearing and three on vision. These items might form the basis for candidate bolt-ons for the EQ-5D.

Eleven items were found to have loadings <0.45 on all components, and these were the SF-6D role, the 15D sexual activities and elimination, the AQoL 8D happiness, contentment with life, social exclusion, communication and enthusiasm and the ICECAP feeling settled and secure, enjoyment and pleasure and achievement and progress. These items might also represent candidate bolt-ons for the EQ-5D.

Generally, items loadings were smaller in EFA compared to PCA. In three cases, and more specifically for AQoL confidence, AQoL burden to others and SF-6D social functioning, smaller loadings implied that the threshold of 0.45 was not met. As a consequence, these three items were identified as candidate bolt-ons in EFA but not in PCA. In other two cases, for 15D breathing and AQoL social isolation, lower loadings meant that the threshold of 0.45 was not met for the cross-loadings on factor not covered by the EQ-5D, resulting in those items being identified as bolt-ons in PCA but not EFA. In one case, for AQoL happiness, the loading was smaller in PCA than in EFA, resulting in the item being identified as a bolt-on with this technique but not with EFA.

In most cases, patterns of relationships between items and component or factors were the same between PCA and EFA. Differences were found for only 4 items and more specifically: AQoL social exclusion and social isolation mainly loaded on relationships in PCA but on psychological symptoms in EFA; SF-6D role loaded on physical functioning in PCA but psychological symptoms in EFA; and HUI 3 dexterity loaded on speech/cognition in PCA but physical functioning in EFA. These differences did not affect the bolt-on identified, but warn on the need of using hypothesis testing techniques to verify the relationship between items and factors/components.

As expected, factors and components were found to correlate. For example, the psychological symptoms component correlated substantially with the satisfaction/contentment, relationships, physical functioning and energy/vitality ones, while the pain component had a strong correlation with the physical functioning construct.

Confirmatory factor analysis

Two 9 factor first order confirmatory models were developed based on PCA results and tested. Model I included items with regression path loadings ≥0.3, while Model II items with regression path loadings ≥0.45. None of the models included the ICECAP achievement and progress item, as this presented loadings <0.3 for all components. Both models showed an acceptable fit, with Model I reporting a RMSEA of 0.056 (90% CI [0.056, 0.057]) and a CFI of 0.924 and Model II a RMSEA of 0.06 (90% CI [0.06, 0.061]) and a CFI of 0.918. Given that Model I fitted data marginally better, this was taken as a starting point for model improvement.

Four items were removed from Model I (15D eating and elimination, AQoL 8D frequency of pleasure and ICECAP Enjoyment and pleasure) as their content was considered theoretically inconsistent with the constructs on which they loaded. Subsequently, the model underwent an iterative process of specification, analysis and respecification. Two items were excluded as they showed local dependencies and large RC with numerous other variables (PWI satisfaction with health and AQoL 8D contentment with life). The final model exhibited good fit, with an RMSEA of 0.041 (90% CI [0.040, 0.041]) and a CFI of 0.963. This was fitted to the second random half of the dataset, reporting once again good fit, with a RMSEA of 0.039 (90% CI [0.039, 0.040]) and a CFI of 0.965. All random resampling of the dataset showed that the model was robust using different subset of the observations. Factor loadings for the final confirmatory model are presented in table 4.4. The factor correlation matrix is presented in table 4.5 and goodness of fit indexes for the random resampling are available in table 4.6.

CFA substantially confirmed the results of categorical PCA and EFA in terms of number of factors identified and items relationship with factors. However, five candidate bolt-ons found with PCA were not confirmed using CFA as they had higher loadings in the CFA (greater than 0.45) compared to PCA. These were the SF-6D role, the 15D sexual activities and the HUI 3 dexterity which all loaded onto physical functioning; and the AQoL 8D happiness and the ICECAP feeling settled which loaded on psychological symptoms. Two additional items, AQoL 8D social isolation and AQoL 8D social exclusion were not confirmed as candidate bolt-ons as their main loading was on relationships in PCA but psychological symptoms in CFA. One item, the 15D breathing, reported a large loading on energy/ sleep in PCA that was not confirmed in CFA.

Eight candidate bolt-ons found with EFA were not confirmed using CFA as they had higher loadings in CFA compared to EFA. These were the SF-6D role, the SF-6D social functioning, the AQoL social social exclusion, the AQoL confidence and the AQoL burden to others and the ICECAP feeling settled and secure which loaded on psychological symptoms; and the HUI 3 dexterity and 15D sexual activities which loaded on physical functioning. Some items presented cross-loadings on one or more factors. If developed into bolt-on dimensions, these items would most likely be sensitive to multiple health aspects.

Structural equation modeling

CFA measurement model was extended into a full structural equation model as described in the methods section. The model exhibited good fit, with a RMSEA of 0.043 (90% CI [0.042, 0.043]) and a CFI of 0.959. Respecifications of the structural paths were considered for those coefficients that were not statistically significant or low (0.05<). One path, relationships regressed on pain, was identified as problematic i.e. very low coefficient. This was eliminated and the model re-estimated. The post hoc model reported an improved fit, with an RMSEA of 0.042 (90% CI [0.041, 0.042]) and a CFI of 0.961. This model was taken as the final model and fitted on the second random half of the dataset. The cross-validated model registered a good fit, with an RMSEA of 0.041 (90% CI [0.040, 0.041]) and a CFI of 0.963. A simplified SEM model without factor correlations is presented in figure 4.1. Table 4.7 reports structural model path estimates and coefficients for factor correlations.

Content analysis showed that EQ-5D dimensions cover constructs related to symptoms and functioning. As a consequence, the first method of SEM selection included only bolt-ons related to these constructs. This led to the selection of 28 candidate bolt-ons across these constructs.

Using the second method of bolt-on selection, 22 candidate bolt-ons were selected. Bolt-ons related to the factors relationships, cognition/speech, energy/vitality, hearing and vision were excluded, as these factors are dependent factors of satisfaction in the validated SEM model. However, as discussed later, the inclusion or exclusion of these candidate items depends on whether satisfaction is selected as a bolt-on. Some items presented cross-loadings on one or more factors. If developed into bolt-on dimensions, these items would most likely be sensitive to multiple health aspects.

Bolt-ons identified using categorical PCA, EFA and CFA are presented in Table 4.8. Bolt-ons selected using SEM are presented in Table 4.9.

One of the additional model structures tested (the model regressed hearing on speech and relationships on psychological symptoms) presented similar goodness of fit (RMSEA of 0.040 (90% CI [0.040, 0.041]) and a CFI of 0.963) to the final structural equation model retained. The model is presented in figure 4.2. This study retained the model presented in figure 4.1 as, after consultation between the supervisory team and an expert in latent structure analysis i.e. Jakob Bue Bjorner, this appeared the most consistent model with the Wilson and Clearly conceptual framework. However, there is no empirical superiority of the retained model compared to the other model structure tested.

4.4 Discussion

This chapter investigated the potential of using PCA, EFA, CFA and SEM for GPBMs bolt-on identification. The results have helped in identifying 9 factors to which GPBMs and SWBMs relate. Of them, 6 potential independent factors and 44 potential independent items may be considered for add-on dimensions to the EQ-5D. Among the items, 37 were identified as their main loading was on a factor not related to the dimensions covered by the EQ-5D and 7 as they were not related to the latent strcture identified.

PCA, EFA and CFA were generally concordant in pinpointing to a common factor structure and similar patterns of association between items and factors, while CFA generated a measurement model that was successfully extended into a full SEM model. Bolt-ons identified were generally consistent across PCA, EFA, CFA and SEM, with only 8 of the 58 bolt-ons found using PCA not being confirmed using CFA and SEM, and similarly only 8 of the 58 bolt-ons found using EFA not being confirmed using CFA and SEM.

Similar results in terms of factor structure and identified bolt-ons might erroneously suggest that these methods can always be used interchangeably. However, this is not the case, as EFA, PCA, CFA and SEM are different techniques that differ in their final goals. While EFA and PCA explore item inter-correlations with the objective of reducing the items into a smaller set of components, CFA and SEM are inferential techniques that test hypothesis on the validity of a predefined measurement model and of a causal structure. The results of this study suggest that there is no variation between the factors identified using PCA, EFA and CFA. By contrast, in some occasions items loading on one factor using PCA and EFA were not confirmed as loading on the same factor using CFA. If the interest is in identifying factors as bolt-on dimensions, PCA, EFA and CFA appear equivalent and interchangeable. However, if the interest is in aidentifying items as bolt-ons, PCA and EFA alone are insufficient for identifying bolt-ons, since it is important to be able to undertake some form of hypothesis testing (e.g. on the impact of adding items, specifying cross-loadings or specifying residual variances). If this latter objective is pursued, it would be advisable to use CFA only after having used exploratory techniques to investigate the latent structure, as suggested in scale development and construct validation studies (e.g. Brown, 2006). This helps accounting for the complexities associated with measuring health constructs. Finally, as SEM specifies latent regression relationships based on a CFA measurement model, CFA is a necessary prerequisite of a SEM study. At the same time, SEM provides some additional information to select bolt-ons when the selection involves multiple candidate items/factors, as it can help to avoid double counting (i.e. if one bolt-on has an impact on another, they should not be added together, as in the case of satisfaction in the current study). However, a prerequisite for a correct use of SEM is having a strong conceptual model to inform the empirical investigation. This study agreed on a conceptual model for the purpose of methodologically investigating the feasibility of using this technique empirically. However, it also highlighted that different models might report similar fits. This suggests a risk of arbitrariness for the selection of bolt-ons using SEM in absence of a universally agreed conceptual model.

The set of strategies employed for identifying bolt-ons were broadly based on how well the EQ-5D covered the factorial structure identified and how strongly items loaded on factors. Criteria used are based on suggested cut-offs in the literature, as well as specifications of the models and reliance on the W&C model. Changes in these choices may result in different bolt-ons being identified. Further consideration of criteria may improve the identification process.

An issue that has not been fully covered in this study is how to select bolt-ons from the identified lists, as not all the items/factors can be added to the EQ-5D (this will be further discussed in Chapter 5 and 6). Albeit affected by substantial conceptual problems, using SEM to select items or factors based on their position in the causal pathways, or on whether they cover relevant parts for a health measure could be one approach. These choices have an impact on what bolt-ons are selected and should therefore be linked to the aim of the bolt-on study, as the aim of developing a broader measure for assessing areas beyond health is different from developing a broader measure of health, and results in different bolt-ons being relevant.

Another selection issue regards the methods for choosing between items related to the same factor. From a theoretical point of view all items loading on a factor not covered by the GPBM are potential candidate bolt-ons. However, more detailed selection procedures might be investigated, an example of which might be choosing items in order of loadings strength. Candidate bolt-ons could also be selected or assessed on the basis of their impact on people's lives or based on the preferences of individuals.

Furthermore, inclusion of bolt-ons may have an impact on the utility values for core items/dimensions of the GPBM. Early studies found interaction effects between bolt-ons and the core dimensions of GPBMs, as well as interaction effects between bolt-ons and the severity of health states, suggesting that simple additive models are likely to lack appropriateness (Brazier et al, 2011). Further research is required to test whether more complex models e.g. multiplicative could be used to establish the impact of bolt-ons on the other dimensions values, or whether full valuations of the bolt-on alongside the original measure is required (Yang et al, 2015).

This study has some limitations. Clustering of items into factors might have been influenced by the health status of the surveys responders e.g. speech/ cognition might have clustered into a factor as responders affected by a chronic condition might report co-occuring problems and not because these items are measuring the same aspects of health. This might reduce the generalizability of the results obtained in different populations. Moreover, disease specific measures were not included in the analysis. As these might describe constructs not covered by GPBMs and SWBMs, some candidate bolt-ons might have been missed. This study therefore suggests a valid method for bolt-on identification only if the dimension/s of interest are already covered by at least one of the measures included in the dataset.

However, also dimensions that have not been included in any existing measure might represent equally relevant candidate bolt-ons. For example, this might be one of the reasons why some categories of the content analysis such as physical symptoms were covered only by one theme i.e. pain. In this latter case different techniques, such as qualitative interviews with patients, might be needed for identifying bolt-ons (see for example the case in mental health in Brazier et al (2014)).

In the analysis, cross-validation was performed on a random half of the dataset, as a second database including the same GPBMs and SWBMs does not exist to date. This might have reduced the statistical power of the analysis. Furthermore, only one strategy for identifying bolt-ons for PCA, EFA and CFA and SEM was employed. As different approaches might determine different bolt-on identification, a full analysis of the impact of identifying bolt-ons using different criteria might have increased the confidence in the results obtained.

Despite these limitations, this study constitutes an important effort in that it is the first study ever conducted to clarify that PCA, EFA, CFA and SEM can be used for identifying bolt-ons in GPBMs, it proposes one strategy for selecting bolt-ons using each of these statistical techniques when the purpose is expanding a GPBMs descriptive system coverage of health constructs, and it highlights the need of using these methods in an integrated manner in order to appropriately select bolt-ons. It is important to highlight that the tables presented in this chapter are the results of a series of steps required for identifying bolt-ons for any possible GPBMs. As a result, these tables could also be used to identify bolt-ons for GPBMs other than the EQ-5D without the need of further analysis. This would involve selecting those factors and items not related to the latent structure measured by the GPBM of interest e.g. vision for the SF-6D.

4.5 Conclusions

PCA, EFA, CFA and SEM can be used for the purpose of identifying bolt-ons. These techniques can be used interchangeably if the objective is identifying factors as bolt-ons. If the interest is also in identifying items to adapt into bolt-on dimensions, both PCA and EFA require the use of CFA for confirming items and factors relation. SEM can be used complementary to CFA to select between factors only if a strong universally agreed conceptual model exists. This chapter has identified 6 factors, and 37 items loading on them, which were not related to the EQ-5D-5L and might for this reason represent candidate bolt-on dimensions for the GPBM. It also identified 7 items not related to the latent strcture. Also these items represent candidate bolt-ons. Further studies applying these techniques on other data sets and different research methods (e.g. qualitative interviews with patients) are needed to inform the selection of bolt-on items.

Table 4.1 – Content analysis based on Wilson and Cleary model

Symptom status			
Physical symptoms	*Measure*	*Psychological symptoms*	*Measure*
Pain		Anxiety/depression	
EQ-5D pain/discomfort	GPBM	EQ-5D anx/depression	GPBM
SF-6D pain	GPBM	15 D depression	GPBM
HUI 3 pain	GPBM	15 D distress	GPBM
15 D discomfort / symptoms	GPBM	AQoL 8D calmness	GPBM
AQoL 8D frequency of pain	GPBM	AQoL 8D depression	GPBM
AQoL 8D intensity of pain	GPBM	AQoL 8D worry	GPBM
AQoL 8D pain inter usual act*	GPBM	ONS anxiety yesterday	SWBM
Other items		Coping	
15 D breathing	GPBM	AQoL 8D worthlessness	GPBM
		AQoL 8D self harm	GPBM
		AQoL 8D feel control life	GPBM
		AQoL 8D freq despair	GPBM
		AQoL 8D coping	GPBM
		Isolation and exclusion	
		AQoL 8D social exc*	GPBM
		AQoL 8D social isol*	GPBM
		Other items	
		SF-6D mental health	GPBM
		AQoL 8D freq sadness	GPBM
		AQoL 8D freq anger	GPBM

		AQoL 8D feeling burden *	GPBM
Functioning status			
Physical functioning	*Measure*	*Psychological functioning*	*Measure*
Independent living/self care		Energy	
EQ-5D self care	GPBM	SF-6D vitality*	GPBM
15 D eating	GPBM	15 D vitality*	GPBM
AQoL 8D get help out house	GPBM	AQoL 8D energy*	GPBM
AQoL 8D need help house job	GPBM		
AQoL 8D washing yourself	GPBM	Other items	
ICECAP being independent	SWBM	HUI 3 cognition*	GPBM
		15 D mental function	GPBM
Ambulation			
EQ-5D mobility	GPBM		
SF-6D physical functioning	GPBM		
HUI 3 ambulation	GPBM		
15 D mobility	GPBM		
AQoL 8D mobility	GPBM		
Senses			
HUI 3 vision	GPBM		
HUI 3 hearing	GPBM		
15 D vision	GPBM		
15 D hearing	GPBM		
AQoL 8D vision	GPBM		
AQoL 8D hearing	GPBM		
Usual activities			

EQ-5D usual activities 15 D usual activities Communication HUI 3 speech 15 D speech AQoL 8D communication Other items 15 D sleeping AQoL 8D sleeping 15 D elimination 15 D sexual activities	GPBM GPBM GPBM GPBM GPBM GPBM GPBM GPBM GPBM		
Social functioning	*Measure*	*Role functioning*	*Measure*
 Relationships AQoL 8D rel (fam and friends) AQoL 8D soc funct (close rel) AQoL 8D enjoy close rel* ICECAP love, friend, support Other items SF-6D social functioning AQoL 8D role in community	 GPBM GPBM GPBM SWBM GPBM GPBM	SF-6D role* AQoL 8D role in the family	GPBM GPBM

General health perception	*Measure*	**Overall quality of life (happiness, satisfaction and subjective wellbeing)**	*Measure*
PWI satisfaction health	SWBM		
		<u>Contentment with life</u>	
		HUI 3 emot (happy life)	GPBM
		AQoL 8D cont life	GPBM
		PWI sat life as a whole	SWBM
		PWI sat standard of living	SWBM
		SWLS life close ideal	SWBM
		SWLS cond life excelent	SWBM
		SWLS satisfaction life	SWBM
		SWLS changing life over	SWBM
		ONS satisfied with life	SWBM
		<u>Purposefulness</u>	
		PWI sat achievement	SWBM
		SWLS got import things	SWBM
		ICECAP achievement	SWBM
		ONS life worthwhile	SWBM
		<u>Happiness</u>	
		AQoL 8D happiness	GPBM
		AQoL 8D pleasure	GPBM
		ICECAP enjoy pleasure	SWBM
		ONS happiness yesterday	SWBM
		<u>Safety</u>	

		PWI sat future security PWI sat safety ICECAP feeling secure <u>Other items</u> PWI sat personal rel* PWI sat feel part comm* PWI sat spirituality	SWBM SWBM SWBM SWBM SWBM SWBM
Characteristic of the individual	*Measure*		
HUI 3 dexterity* AQoL 8D confid in yourself AQoL 8D enthusiam	GPBM GPBM GPBM		

Note: * items that theoretically relate to more than one category of the Wilson and Cleary model. Cross-loadings might be expected in PCA and EFA. comm= community; cond= condition; confid=confidence; content= contentment; emot= emotion; enjoy= enjoyment; exc= exclusion; fam friends= family and friends; feel= feeling; freq=frequency; import = important; get help out house= getting help outside house; inter usual act=interference usual activities; isol= isolation; job= jobs; rel = relationships; sat= satisfaction; soc func= social functioning.

Table 4.2 – Categorical PCA pattern matrix with item loadings ≥ 0.3 for a 9-component solution with Promax rotation

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
EQ-5D mobility	**.887**								
AQoL 8D mobility	**.872**								
HUI 3 ambulation	**.865**								
15 D mobility	**.857**								
AQoL 8D get outside house	**.830**								
AQoL 8D get help house jobs	**.817**								
SF-6D physical functioning	**.797**								
EQ-5D usual activities	**.796**								
AQoL 8D washing yourself	**.774**								
EQ-5D self-care	**.755**								
15 D usual activities	**.740**								
AQoL 8D role community	**.710**								

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
AQoL 8D role family	**.576**								
ICECAP being independent	**.515**								
15 D breathing✓	**.474**							**.457**	
SF-6D role✓	.412	.363							
15 D sexual activities ✓	.356							.315	
SF-6D mental health		**.963**							
EQ-5D anxiety/depression		**.909**							
AQoL 8D feeling of despair		**.892**							
15 D distress		**.889**							
AQoL 8D frequency worry		**.854**							
AQoL 8D frequency depression		**.840**							
AQoL 8D sadness		**.830**							
15 D depression		**.806**							

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
AQoL 8D frequency worthlessness		**.734**							
ONS Anxiety yesterday		**.712**			-.300				
AQoL 8D calmness or agitation		**.656**							
AQoL 8D anger		**.608**							
AQoL 8D self harm		**.591**							
AQoL 8D coping with life problems		**.578**							
AQoL 8D burden to others	.393	**.529**							
SF-6D social functioning	**.489**	**.528**							
AQoL 8D feeling in control of life		**.514**							
AQoL 8D confidence		**.508**							
HUI 3 emotion		**.500**							
AQoL 8D happiness ✓		.439			.350				
ICECAP Feeling settled and secure ✓		.366							

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
AQoL 8D contentment with life ✓		.339							
PWI satisfaction standard of living ✓			**.899**						
ONS satisfaction life ✓			**.889**						
PWI satisfaction life as a whole ✓			**.858**						
PWI satisfaction future security ✓			**.829**						
PWI satisfaction achievement ✓			**.826**						
SWLS satisfaction with life ✓			**.825**						
SWLS conditions life excellent ✓			**.823**						
SWLS life close to ideal ✓			**.779**						
SWLS gotten important things in life ✓			**.776**						
ONS things you do are worthwhile ✓			**.747**						
PWI satisfaction with safety ✓			**.736**						
ONS happiness yesterday ✓			**.681**						

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
PWI satisfaction community ✓			**.660**						
SWLS changing life over ✓			**.614**						
PWI satisfaction personal relationships ✓			**.581**		**-.487**				
PWI satisfaction with spirituality ✓			**.567**						
PWI satisfaction with health ✓	-.304		**.512**						
AQoL 8D intensity pain/discomfort				**.759**					
EQ-5D pain/discomfort				**.742**					
HUI 3 pain	.322			**.700**					
AQoL 8D frequency serious pain				**.689**					
15 D discomfort and symptoms				**.658**					
AQoL 8D pain interference with usual	.392			**.559**					
SF-6D Pain	.414			**.515**					
AQoL 8D close rel (including sexual) ✓					**.896**				

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
ICECAP love, friendship and support ✓					**.831**				
AQoL 8D close relationships (family) ✓					**.800**				
AQoL 8D enjoyment close relationship ✓					**.680**				
AQoL 8D frequency pleasure ✓					**.501**				
AQoL 8D social isolation ✓		.421			**.453**				
AQoL 8D social exclusion ✓		.408			.448				
ICECAP enjoyment and pleasure ✓					.445				
15 D speech ✓						**.763**			
15 D eating ✓						**.693**			
HUI 3 speech ✓						**.677**			
15D mental function ✓						**.535**		**.514**	
HUI 3 cognition ✓						**.474**		.434	
HUI 3 dexterity ✓	.304					**.461**			

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
AQoL 8D communication ✓						.449			
15 D hearing ✓							**.869**		
AQoL 8D hearing ✓							**.852**		
HUI 3 hearing ✓							**.815**		
AQoL 8D energy ✓								**.688**	
SF-6D vitality ✓	.302							**.600**	
AQoL 8D sleeping ✓				.388				**.524**	
15 D sleeping ✓				.398				**.504**	
15 D vitality ✓								**.477**	
AQoL 8D enthusiasm ✓								.440	
15 D elimination ✓								.412	
AQoL 8D vision ✓									**.911**
15D vision ✓									**.863**

	Components								
	Physical functioning	Psycholog symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing✓	Energy / Sleep✓	Vision✓
HUI 3 vision ✓									**.772**
ICECAP achievement and progress✓									

Note: Loadings in bold are considered relevant, good or very good following Comrey and Lee classification. Red refers to items with loadings<0.45 on all constructs.

Psycholog sympt=psychological symptoms;

✓ Represent components or individual items identified as candidate bolt-ons for the EQ-5D using categorical PCA.

Table 4.3 – EFA pattern matrix with item loadings ≥ 0.3 for a 9-factor solution with Promax rotation

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
EQ-5D mobility	**.952**								
HUI 3 ambulation	**.944**								
15 D mobility	**.910**								
AQoL 8D mobility	**.875**								
AQoL 8D washing yourself	**.805**								
EQ-5D Usual activities	**.796**								
AQoL 8D get help house jobs	**.795**								
AQoL 8D get around outside house	**.792**								
15 D usual activities	**.782**								
EQ-5D Self-care	**.780**								
SF-6D Physical functioning	**.768**								
AQoL 8D role in your community	**.644**								

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
AQoL 8D role in the family	**.530**								
15 D breathing	**.470**								
ICECAP being independent	**.469**								
15 D sexual activities ✓	.430								
HUI 3 dexterity ✓	.394								
15 D distress		**1.002**							
EQ-5D anxiety/depression		**.978**							
SF-6D mental health		**.976**							
AQoL 8D feeling of despair		**.922**							
AQoL 8D frequency worry		**.910**							
AQoL 8D frequency depression		**.908**							
15 D depression		**.902**							
AQoL 8D sadness		**.819**							

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
ONS Anxiety yesterday		**.691**							
AQoL 8D frequency worthlessness		**.672**							
AQoL 8D calmness or agitation		**.625**							
HUI 3 emotion		**.553**							
AQoL 8D anger		**.534**							
AQoL 8D coping with life problems		**.532**							
AQoL 8D self harm		**.509**							
AQoL 8D feeling in control of life		**.482**							
AQoL 8D happiness		**.459**							
AQoL 8D social isolation		**.451**							
AQoL 8D burden to others ✓	.356	.437							
SF-6D social functioning ✓	.421	.439							
AQoL 8D confidence ✓		.403						.395	

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
SF-6D role ✓	.297	.444							
AQoL 8D social exclusion ✓		.423							
ICECAP Feeling settled and secure ✓		.403							
AQoL 8D contentment with life ✓		.353	.364						
PWI satisfaction standard of living ✓			**-.892**						
ONS satisfaction life ✓			**-.886**						
SWLS satisfaction with life ✓			**-.882**						
PWI satisfaction achievement ✓			**-.870**						
PWI satisfaction life as a whole ✓			**-.857**						
SWLS conditions life excellent ✓			**-.839**						
SWLS life close to ideal ✓			**-.836**						
PWI satisfaction future security ✓			**-.796**						
SWLS gotten important things in life ✓			**-.768**						

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
ONS things you do are worthwhile ✓			**-.679**						
PWI satisfaction with safety ✓			**-.638**						
PWI satisfaction personal relations ✓			**-.605**		-.470				
ONS happiness yesterday ✓		-.367	**-.579**						
PWI satisfaction community ✓			**-.591**						
SWLS changing life over ✓			**-.591**						
PWI satisfaction with health ✓			**-.485**						
PWI satisfaction with spirituality ✓			-.400						
EQ-5D pain/discomfort	.370			**.711**					
AQoL 8D intensity pain/discomfort				**.709**					
HUI 3 pain	.411			**.649**					
15 D discomfort and symptoms	376			**.595**					
AQoL 8D frequency serious pain	.385			**.578**					

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
SF-6D Pain	.433			**.530**					
AQoL 8D pain interference with usual	.476			**.487**					
AQoL 8D enjoyment close relationship ✓					**.745**				
AQoL 8D close relationships (family) ✓					**.714**				
ICECAP love, friendship and support ✓					**.699**				
AQoL 8D close relationships (including sexual) ✓					**.559**				
AQoL 8D frequency pleasure ✓					.315				
ICECAP enjoyment and pleasure ✓					.313				
15 D speech ✓						**.748**			
HUI 3 speech ✓						**.705**			
15 D eating ✓						**.512**			

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
15D mental function ✓						**.493**		.296	
AQoL 8D communication ✓						**.450**			
HUI 3 cognition ✓						.434		.328	
15 D hearing ✓							**.859**		
AQoL 8D hearing ✓							**.711**		
HUI 3 hearing ✓							**.681**		
AQoL 8D energy ✓								**.723**	
SF-6D vitality ✓								**.630**	
AQoL 8D enthusiasm ✓								**.530**	
15 D vitality ✓		.313						.413	
AQoL 8D sleeping ✓								.361	
ICECAP achievement and progress ✓								.337	
15 D sleeping ✓								.327	

	Factors								
	Physical functioning	Psychological symptoms	Satisfaction ✓	Pain	Relationships ✓	Speech / Cognition ✓	Hearing ✓	Energy / Sleep ✓	Vision ✓
AQoL 8D vision ✓									**.671**
15D vision ✓									**.672**
HUI 3 vision ✓									**.535**
15D elimination ✓									

Note: Loadings in bold are considered relevant, good or very good following Comrey and Lee classification. Red refers to items with loadings<0.45 on all constructs.

✓Represent factors or individual items identified as candidate bolt-ons for the EQ-5D using standard exploratory factor analysis

Table 4.4 – CFA standardized factor loadings and standard errors

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
EQ-5D usual activities	0.931 (0.005)								
AQoL 8D get around outside house	0.906 (0.005)								
AQoL 8D get help house jobs	0.883 (0.006)								
AQoL 8D washing yourself	0.878 (0.008)								
EQ-5D Self-care	0.875 (0.011)								
AQoL 8D role in your community	0.692 (0.014)	0.149 (0.031)	0.102 (0.022)		0.185 (0.033)				
EQ-5D mobility	0.873 (0.007)								
AQoL 8D role in the family	0.579 (0.014)	0.226 (0.029)	0.108 (0.022)		0.235 (0.032)				

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
AQoL 8D mobility	0.857 (0.007)								
15 D mobility	0.831 (0.009)								
HUI 3 ambulation	0.828 (0.010)								
SF-6D Physical functioning	0.817 (0.007)								
15 D sexual activities	0.746 (0.012)								
15 D breathing	0.667 (0.013)								
ICECAP being independent	0.698 (0.014)								
AQoL 8D dexterity	0.653 (0.019)								

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
SF-6D role	0.485 (0.015)	0.351 (0.016)							
SF-6D social functioning	0.449 (0.014)	0.433 (0.014)							
AQoL 8D pain interference with usual	0.286 (0.018)			0.667 (0.017)					
15 D depression		0.929 (0.004)							
AQoL 8D frequency depression		0.911 (0.004)							
AQoL 8D happiness		0.899 (0.005)							
EQ-5D anxiety/depression		0.896 (0.005)							
AQoL 8D sadness		0.893 (0.004)							

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
AQoL 8D frequency worthlessness		0.873 (0.006)							
AQoL 8D feeling of despair		0.866 (0.007)							
AQoL 8D feeling in control of life		0.856 (0.006)							
AQoL 8D coping with life problems		0.848 (0.007)							
15 D distress		0.844 (0.007)							
SF-6D mental health		0.797 (0.007)							
AQoL 8D frequency worry		0.793 (0.007)							
AQoL 8D confidence		0.760 (0.008)							

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
AQoL 8D calmness or agitation		0.744 (0.009)							
HUI 3 emotion		0.703 (0.012)	-0.278 (0.013)						
AQoL 8D self harm		0.704 (0.014)							
ICECAP Feeling settled and secure		0.623 (0.016)	-0.257 (0.017)						
AQoL 8D anger		0.593 (0.012)							
AQoL 8D burden to others	0.350 (0.016)	0.529 (0.015)							
AQoL 8D social isolation		0.515 (0.022)			0.332 (0.024)				
AQoL 8D social exclusion		0.503 (0.025)			0.295 (0.026)				

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
ONS Anxiety yesterday		0.400 (0.016)							
PWI satisfaction standard of living ✓			0.709 (0.012)						
PWI satisfaction achievement ✓			0.806 (0.009)						
ONS satisfaction life ✓			0.904 (0.007)						
PWI satisfaction life as a whole ✓			0.840 (0.008)						
SWLS satisfaction with life ✓			0.923 (0.005)						
SWLS conditions life excellent ✓			0.868 (0.007)						
SWLS life close to ideal ✓			0.866 (0.008)						

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
SWLS gotten important things in life ✓			0.777 (0.008)						
ONS things you do are worthwhile ✓			0.824 (0.009)						
SWLS changing life over ✓			0.700 (0.011)						
ONS happiness yesterday ✓			0.830 (0.010)						
PWI satisfaction personal relations ✓			0.652 (0.013)		0.025 (0.027)				
PWI satisfaction community ✓			0.642 (0.013)						
PWI satisfaction with safety ✓			0.634 (0.013)						
PWI satisfaction future security ✓			0.725 (0.012)						

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
PWI satisfaction with spirituality ✓			0.410 (0.016)						
AQoL 8D intensity pain/discomfort				0.942 (0.004)					
EQ-5D pain/discomfort				0.943 (0.003)					
HUI 3 pain				0.950 (0.003)					
AQoL 8D frequency serious pain				0.914 (0.005)					
15 D discomfort and symptoms				0.929 (0.004)					
SF-6D Pain				0.925 (0.004)					
AQoL 8D enjoyment close relationship ✓					0.715 (0.011)				
ICECAP Love, friendship and support ✓					0.787 (0.010)				

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
AQoL 8D close relationships (family) ✓					0.810 (0.009)				
AQoL 8D close rel (including sexual)✓					0.782 (0.009)				
AQoL 8D hearing ✓						0.956 (0.013)			
15 D hearing ✓						0.934 (0.013)			
HUI 3 hearing ✓						0.821 (0.015)			
AQoL 8D communication ✓						-0.049 (0.047)*	0.869 (0.043)		
15 D speech ✓							0.899 (0.021)		
HUI 3 speech ✓							0.852 (0.020)		
15D mental function✓							0.485 (0.022)		0.415 (0.022)

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
HUI 3 cognition ✓							0.440 (0.024)		0.392 (0.022)
15D vision ✓								0.918 (0.020)	
AQoL 8D vision ✓								0.815 (0.016)	
HUI 3 vision ✓								0.540 (0.025)	
15 D vitality ✓									0.925 (0.005)
AQoL 8D enthusiasm ✓		0.297 (0.024)	-0.071 (0.017)		0.298 (0.021)				0.222 (0.023)
AQoL 8D energy ✓									0.853 (0.006)
SF-6D vitality ✓									0.796 (0.009)

	Factors								
	Physical functioning	Psychogical symptoms	Satisfaction ✓	Pain	Relationships ✓	Hearing ✓	Speech / Cognition ✓	Vision ✓	Energy / Sleep ✓
15 D sleeping ✓									0.696 (0.011)
AQoL 8D sleeping ✓									0.676 (0.011)

Note: * loadings not statistically significant at 0.01 level; ✓identified bolt-ons using CFA.

Table 4.5 – CFA factor correlation matrix estimates (standard errors)

Component	*Physical functioning*	*Psychological symptoms*	*Satisfaction*	*Pain*	*Relationships*	*Speech /Cognition*	*Hearing*	*Energy / Sleep*	*Vision*
Physical functioning	**1.000**	0.495 (0.014)	-0.365 (0.016)	0.836 (0.008)	0.438 (0.016)	0.473 (0.023)	0.334 (0.020)	0.750 (0.009)	0.479 (0.018)
Psychological symptoms	0.495 (0.014)	**1.000**	-0.724 (0.008)	0.428 (0.015)	0.827 (0.007)	0.522 (0.021)	0.161 (0.019)	0.796 (0.007)	0.298 (0.019)
Satisfaction	-0.365 (0.016)	-0.724 (0.008)	**1.000**	-0.317 (0.016)	-0.739 (0.009)	-0.302 (0.023)	-0.105 (0.020)	-0.613 (0.012)	-0.237 (0.019)
Pain	0.836 (0.008)	0.428 (0.015)	-0.317 (0.016)	**1.000**	0.353 (0.016)	0.342 (0.024)	0.268 (0.019)	0.695 (0.010)	0.393 (0.018)
Relationships	0.438 (0.016)	0.827 (0.007)	-0.739 (0.009)	0.353 (0.016)	**1.000**	0.460 (0.022)	0.207 (0.020)	0.697 (0.011)	0.307 (0.019)
Speech / Cognition	0.473 (0.023)	0.522 (0.021)	-0.302 (0.023)	0.342 (0.024)	0.460 (0.022)	**1.000**	0.586 (0.024)	0.421 (0.023)	0.458 (0.025)
Hearing	0.334 (0.020)	0.161 (0.019)	-0.105 (0.020)	0.268 (0.019)	0.207 (0.020)	0.586 (0.024)	**1.000**	0.245 (0.020)	0.465 (0.019)
Energy / Sleep	0.750 (0.009)	0.796 (0.007)	-0.613 (0.012)	0.695 (0.010)	0.697 (0.011)	0.421 (0.023)	0.245 (0.020)	**1.000**	0.406 (0.018)
Vision	0.479 (0.018)	0.298 (0.019)	-0.237 (0.019)	0.307 (0.019)	0.307 (0.019)	0.458 (0.025)	0.465 (0.019)	0.406 (0.018)	**1.000**

Table 4.6 – Goodness of fit indexes for CFA for resampling of the dataset

Resampling		*RMSEA*		*CFI*
Series	*Percentage of the dataset resampled*	*Estimate*	*90% confidence intervals*	
Series 1	50 % of cases	0.041	0.040 – 0.041	0.961
Series 2	50 % of cases	0.039	0.038 – 0.039	0.967
Series 3	50 % of cases	0.040	0.039 – 0.040	0.964
Series 4	50 % of cases	0.039	0.039 – 0.040	0.965
Series 5	50 % of cases	0.041	0.040 – 0.041	0.961
Series 6	50 % of cases	0.040	0.039 – 0.040	0.964
Series 7	50 % of cases	0.040	0.039 – 0.040	0.965
Series 8	50 % of cases	0.040	0.040 – 0.041	0.962
Series 9	50 % of cases	0.041	0.040 – 0.041	0.961
Series 10	50 % of cases	0.039	0.038 – 0.039	0.967
Series 11	10 % of cases	0.036	0.035 – 0.036	0.971
Series 12	20 % of cases	0.038	0.037 – 0.038	0.967
Series 13	30 % of cases	0.039	0.038 – 0.040	0.966
Series 14	40 % of cases	0.040	0.039 – 0.040	0.965

Figure 4.1 – Structural equation model regression paths standardized estimates and standard errors

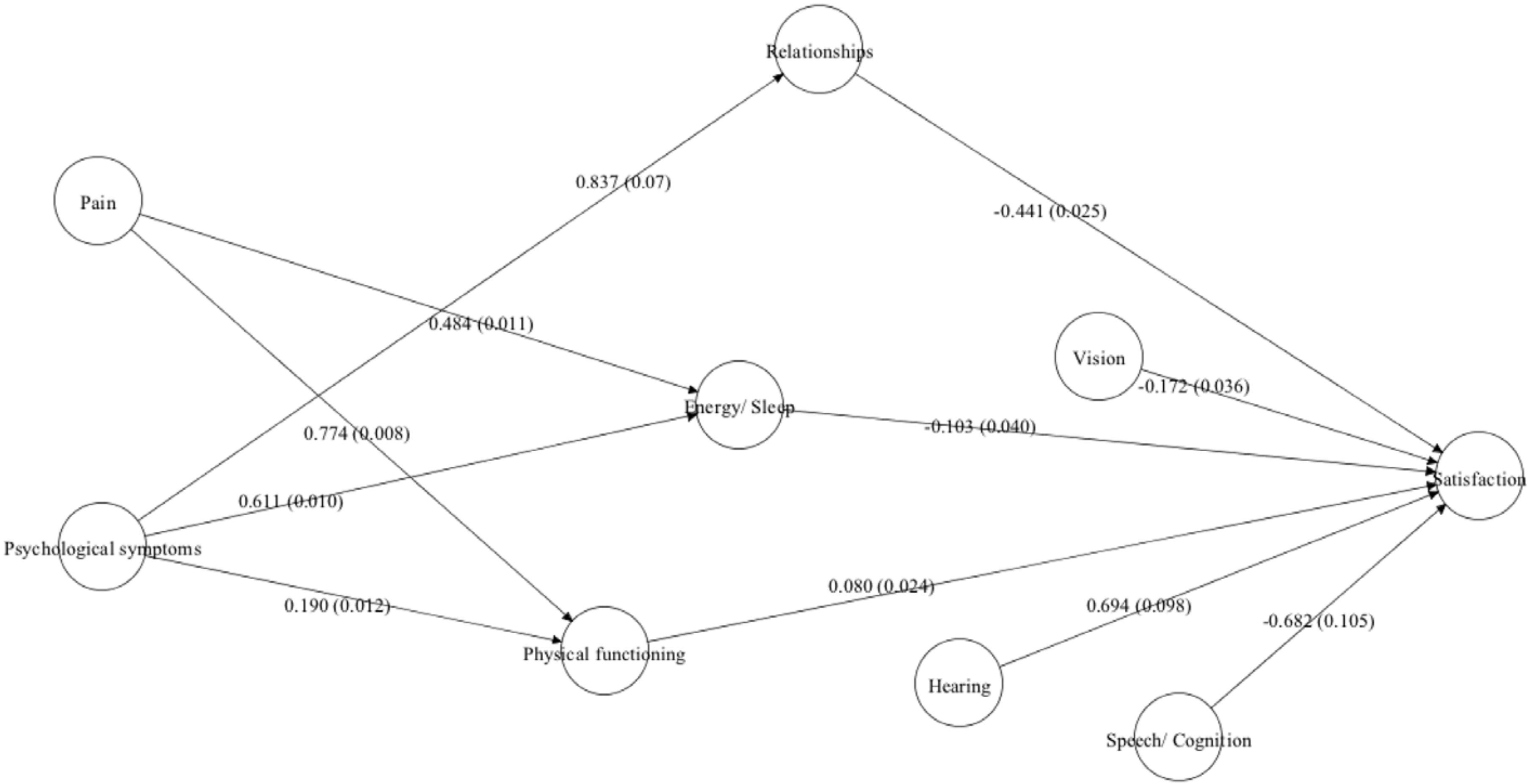


Table 4.7 – Regressions and correlations coefficients for the structural equation model and standard errors

Regressions			
Dependent factor	***Independent factor***	***Estimate (SE)***	***P value***
Relationships	Psychological symptoms	0.837 (0.07)	0.000
Physical functioning	Psychological symptoms	0.190 (0.012)	0.000
	Pain	0.774 (0.008)	0.000
Energy / Sleep	Psychological symptoms	0.611 (0.010)	0.000
	Pain	0.484 (0.011)	0.000
Satisfaction	Relationships	-0.441 (0.025)	0.000
	Physical functioning	0.080 (0.024)	0.001
	Energy / Sleep	-0.103 (0.040)	0.010
	Speech	-0.682 (0.105)	0.000
	Vision	-0.172 (0.036)	0.000
	Hearing	0.694 (0.098)	0.000

Correlations			
Factors		***Estimate (SE)***	***P value***
Psychological symptoms	Pain	0.414 (0.015)	0.000
	Hearing	0.184 (0.019)	0.000
	Vision	0.321 (0.018)	0.000
	Speaking	0.564 (0.024)	0.000
Hearing	Pain	0.329 (0.020)	0.000
	Vision	0.511 (0.019)	0.000
	Speaking	0.899 (0.017)	0.000
Pain	Vision	0.469 (0.018)	0.000
	Speaking	0.448 (0.024)	0.000
Speaking	Vision	0.436 (0.028)	0.000

Table 4.8 – Categorical PCA, EFA and CFA Bolt-ons

	Original measure	***Categorical PCA Bolt-ons***	***Normal EFA***	***CFA Bolt-ons***
Factor / Components		Relationships Cognition / Speech Energy/Sleep Hearing Satisfaction Vision	Relationships Cognition / Speech Energy/Sleep Hearing Satisfaction Vision	Relationships Cognition / Speech Energy/Sleep Hearing Satisfaction Vision
GPBMs	SF-6D	SF-6D role SF-6D vitality	SF-6D role SF-6D social functioning * SF-6D vitality	SF-6D vitality
	HUI 3	HUI 3 cognition HUI 3 dexterity HUI 3 hearing HUI 3 speech HUI 3 vision	HUI 3 cognition HUI 3 dexterity HUI 3 hearing HUI 3 speech HUI 3 vision	HUI 3 cognition HUI 3 hearing HUI 3 speech HUI 3 vision

	15 D	15 D breathing*		
		15 D eating	15 D eating	15 D eating
		15 D elimination	15 D elimination	15 D elimination
		15 D hearing	15 D hearing	15 D hearing
		15D mental function	15D mental function	15D mental function
		15 D sexual activities	15 D sexual activities	
		15 D sleeping	15 D sleeping	15 D sleeping
		15 D speech	15 D speech	15 D speech
		15D vision	15D vision	15D vision
		15 D vitality	15 D vitality	15 D vitality
	AQoL 8D	AQoL 8D close relationships (family)	AQoL 8D close relationships (family)	AQoL 8D close relationships (family)
		AQoL 8D communication	AQoL 8D communication	AQoL 8D communication
		AQoL 8D contentment with life	AQoL 8D contentment with life	AQoL 8D contentment with life
		AQoL 8D energy	AQoL 8D energy	AQoL 8D energy
		AQoL 8D enjoyment close relationship	AQoL 8D enjoyment close relationship	AQoL 8D enjoyment close relationship

		AQoL 8D enthusiasm AQoL 8D frequency pleasure AQoL 8D happiness* AQoL 8D hearing AQoL 8D sleeping AQoL 8D social exclusion AQoL 8D close rel (including sexual) AQoL 8D social isolation* AQoL 8D vision	AQoL 8D enthusiasm AQoL 8D frequency pleasure AQoL 8D hearing AQoL 8D sleeping AQoL 8D social exclusion AQoL 8D close rel (including sexual) AQoL 8D vision AQoL 8D confidence * AQoL 8D burden to others *	AQoL 8D enthusiasm AQoL 8D frequency pleasure AQoL 8D hearing AQoL 8D sleeping AQoL 8D close rel (including sexual) AQoL 8D vision
SWBMs	PWI	PWI satisfaction achievement PWI satisfaction community PWI satisfaction future security	PWI satisfaction achievement PWI satisfaction community PWI satisfaction future security	PWI satisfaction achievement PWI satisfaction community PWI satisfaction future security

		PWI satisfaction life as a whole PWI satisfaction personal relationships PWI satisfaction standard of living PWI satisfaction with health PWI satisfaction with safety PWI satisfaction with spirituality	PWI satisfaction life as a whole PWI satisfaction personal relationships PWI satisfaction standard of living PWI satisfaction with health PWI satisfaction with safety PWI satisfaction with spirituality	PWI satisfaction life as a whole PWI satisfaction personal relationships PWI satisfaction standard of living PWI satisfaction with health PWI satisfaction with safety PWI satisfaction with spirituality
	SWLS	SWLS changing life over SWLS conditions life excellent SWLS gotten important things in life SWLS life close to ideal SWLS satisfaction with life	SWLS changing life over SWLS conditions life excellent SWLS gotten important things in life SWLS life close to ideal SWLS satisfaction with life	SWLS changing life over SWLS conditions life excellent SWLS gotten important things in life SWLS life close to ideal SWLS satisfaction with life
	ICECAP	ICECAP achievement and progress ICECAP enjoyment and	ICECAP achievement and progress ICECAP enjoyment and	ICECAP achievement and progress ICECAP enjoyment and

		pleasure ICECAP feeling settled and secure ICECAP love, friendship and support	pleasure ICECAP feeling settled and secure ICECAP love, friendship and support	pleasure ICECAP love, friendship and support
	ONS	ONS happiness yesterday ONS satisfaction life ONS things you do are worthwhile	ONS happiness yesterday ONS satisfaction life ONS things you do are worthwhile	ONS happiness yesterday ONS satisfaction life ONS things you do are worthwhile

Note: * represents bolt-ons that differ between categorical PCA and EFA. Red represents bolt-ons that differ between categorical PCA and CFA, and between EFA and CFA. Blue represents bolt-ons identified as not related to the factorial structure.

Table 4.9 –SEM Bolt-ons

	Original measure	***SEM Bolt-ons Method I***	***SEM Bolt-ons Method II***
Factor / Components		Relationships Cognition / speech Energy/Sleep Hearing Vision	Satisfaction
GPBMs	SF-6D	SF-6D vitality	
	HUI 3	HUI 3 cognition HUI 3 emotion HUI 3 hearing HUI 3 speech HUI 3 vision	HUI 3 emotion
	15 D	15 D hearing 15D mental function	

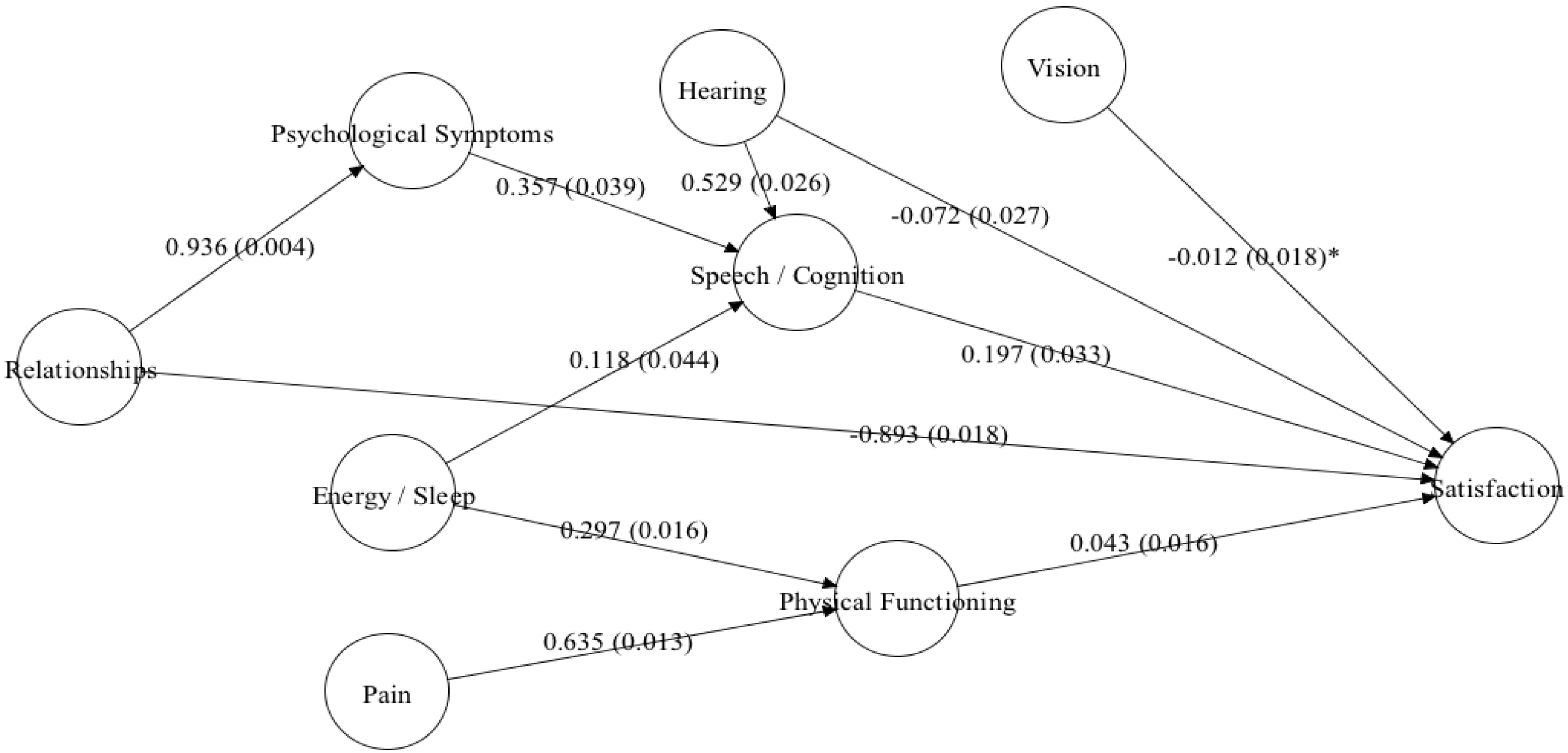
		15 D sleeping 15 D speech 15D vision 15 D vitality	
	AQoL 8D	AQoL 8D close relationships (family) AQoL 8D communication AQoL 8D contentment with life AQoL 8D energy AQoL 8D enjoyment close relationship AQoL 8D frequency pleasure AQoL 8D hearing AQoL 8D sleeping AQoL 8D close relationships (including sexual)	AQoL 8D enthusiasm AQoL 8D role in the community AQoL 8D role in the family

		AQoL 8D vision	
SWBMs	PWI		PWI satisfaction achievement PWI satisfaction community PWI satisfaction future security PWI satisfaction life as a whole PWI satisfaction personal relationships PWI satisfaction standard of living PWI satisfaction with health PWI satisfaction with safety PWI satisfaction with spirituality
	SWLS		SWLS changing life over SWLS conditions life excellent SWLS gotten important things in life

			SWLS life close to ideal SWLS satisfaction with life
	ICECAP	ICECAP love, friendship and support	
	ONS		ONS happiness yesterday ONS satisfaction life ONS things you do are worthwhile

Note: SEM Bolt-ons are a subset of CFA ones

Figure 4.2 – Alternative structural equation model regression paths standardized estimates and standard errors



Chapter 5
Selecting bolt-ons using their ability to predict health related quality of life

5.1 Introduction

Chapter 4 presented a methodological study on the possibility of identifying bolt-ons using PCA, EFA and CFA. The chapter has shown that PCA, EFA and CFA are generally concordant in pinpointing to a common latent structure and similar patterns of association between items and factors, and that the three techniques identify generally consistent bolt-ons. Despite these similarities, in line with other factor analysis literature, CFA should be considered the preferred approach, as hypothesis on the items behavior need to be tested after exploratory work has been conducted for correctly identifying items as bolt-ons. The chapter concluded by presenting 6 factors and 46 items that might represent candidate bolt-on dimensions, 37 of which were identified as their main loading was on a factor not related to the latent structure of the EQ-5D-5L.

Chapter 4 has also introduced the problem of double counting and investigated the possibility of using SEM to account for this. However, double counting is not the only reason why it is necessary to select between the identified items and factors. Evidence to date suggests that the impact of bolt-ons is not simply additive as also coefficients for the core 5 dimensions of the EQ-5D are altered when adding a bolt-on (Brazier et al, 2012). This implies that for each bolt-on option, the new descriptive systems need to undergo a complete re-evaluation to obtain utility values for the health states they describe (see Chapter 3). This process can be costly and complex. Hence, additional methods of selection might be needed even if the identified factors and items measure distinct constructs relevant to expand the investigated GPBM descriptive system.

This chapter investigates whether bolt-ons identified using factor analysis methods can be selected based on their ability to predict differences in HRQoL. This is an important aspect as the dimensions added to the GPBM descriptive system should tap on constructs that are relevant for the HRQoL of patients and the general population. It uses linear regressions, a technique that was chosen as it is extremely common and easy to conduct (Field, 2009). This maximizes the utility of this research for future applications.

5.2 Methods

5.2.1 Data

The regression study presented here used the MIC dataset, the same database used for the factor analysis presented in Chapter 4. Details on the data collection methods can be obtained from Richardson et al (2012 a) and Richardson et al (2012 b), and are described more in depth in Chapter 4. Broadly, an online survey was carried out in six countries (Australia, Canada, Germany, Norway, United Kingdom and United States) imposing quotas to obtain similar socio-demographic characteristics across countries. Responders were members of the general public who had previously agreed to participate to online surveys. The final sample comprised 8022 individuals, 6262 of whom reported to be affected by one of the following 9 chronic health conditions: asthma, cancer, COPD, depression, diabetes, hearing problems, arthritis, heart dieseases and stroke. The remaining 1760 individuals did not report to be affected by any chronic health condition.

5.2.1 Questionnaires, items and factors

The current study used the 5L version of the EQ-5D. In addition, a selection of items from the SF-6D, the HUI3, the AQoL 8D, the 15D, the PWI, the SWLS, the ONS and the ICECAP were employed. The items selected were the 37 identified as potential bolt-ons related to factors not already covered by the EQ-5D presented in Chapter 4. This choice was made to allow comparability of results betwen items and factors regressions. All items from the GPBMs and SWBMs were ordinal categorical, with categories varying between 4 and 11. Items for the PWI, SWBMs and ONS were recoded in order for them to report decrements in HRQoL at decreasing levels of satisfaction (i.e. level 1 perfect satisfaction). The EQ-5D and selected items from the SF-6D, HUI3, AQoL 8D, 15D, PWI, SWLS, ONS and ICECAP were used as independent variables. Wording for the items tested is presented in Table 5.1.

This study used the Health VAS as a dependent variable. This was intended as a proxy of HRQoL. The Health VAS asked responders to give a rating for their health today using 100 and 0 as anchors for excellent health and death, respectively. Excellent health was defined as excellent physical, mental and social health. Physical health was defined as no pain, discomfort or itching, perfect hearing vision and speech, excellent strength, flexibility, movement and energy. Mental health was defined as very happy, enthusiastic and contented, never sad or depressed, confident and with high self worth. Social health was defined as excellent social and family relationships. As the Health VAS reported an approximately normal distribution, this was considered as a continuous variable. Finally, the 6 latent factors reported in the previous chapter and identified in Finch et al (2017), namely satisfaction, hearing, vision, energy/sleep, relationships and speech cognition, were employed. Latent factors are continuous variables and were used as independent variables.

5.2.2 Analyses

Two tests were performed. The first test was carried out to discriminate between bolt-ons in terms of their ability to detect variations in HRQoL not already accounted for by the EQ-5D-5L. The second test further examined whether detected variations helped explaining differences in HRQoL between patients and the general population in 9 chronic conditions. These information may be used to select between factor and items as for bolt-ons to be relevant they should be able to detect aspects of HRQoL not already covered by the parent measure i.e. EQ-5D, and these aspects should help explaining differences in HRQoL between patients and the general population for one or more conditions. The ability of factors and items to detect differeneces in HRQoL might suggest that their addition would improve the validity and responsiveness of the EQ-5D.

Test 1

In order to assess whether different factors and items were able to detect differences in HRQoL as detected by the Health VAS, a base model was estimated regressing the Health VAS over the EQ-5D-5L dummies and socio-demographic controls. The model was subsequently extended with the inclusion of factors and items identified from Chapter 4, each of which was added individually. Unstandardized β coefficients for factors and items are reported. For factors, these indicated the amount of decrease in HRQoL as a result of a unit change in the latent factor tested. For items, these indicated the amount of decrease in HRQoL associated with the level of the dummy variable compared to the reference case (best possible health/ satisfaction). The size, direction and statistical significance of the β coefficients was used to compare factors and items. Comparatively larger β coefficients meant that the factor / item was better in predicting differences in HRQoL not already captured by the EQ-5D-5L. Non statistically significant β coefficients suggested no impact of the factor or item in predicting difference in HRQoL. If the addition of a factor or item made one or more of the EQ-5D-5L dimensions not statistically significant, this meant that it was able to take full account of variations of HRQoL for those dimensions. If the factor or item substantially reduced the coefficient of one or more of the EQ-5D-5L dimensions, this showed a possible interaction between that factor or item and the dimension for which the coefficient was reduced. These latter two pieces of information may be used for selecting factors or items e.g. if a choice between two items related to the same factor has to be made, the item having less impact on the remaining dimensions of the EQ-5D should be chosen as this shows that it has less overlap with the aspects of health already covered by the questionnaire.

To enhance clarity of reporting and comparison of results, a maximum of 4 items per factor are presented for the items regressions. These were selected among those that reported the most prominent results. For each of these items only a mild, a moderate and a severe state were chosen among the one covered. These were selected based on the overall number of levels presented in the item. Full results for all items and all levels of the items are provided in the Appendices Chapter 5 – Table I.

Analyses of observable variables (item regressions) were conducted in STATA/MP 14 ©. Analyses of latent variables (factor regressions) were performed in Mplus version 7©.

Test 2

In order to assess whether different factors and items were able to explain differences in HRQoL between patients and the general population, the second test replicated the methods employed by Bockerman et al (2011). The Health VAS was firstly regressed upon socio-demographic characteristics and dummy variables for asthma, cancer, chronic obstructive pulmonary disease, depression, diabetes, hearing problems, arthritis, heart diseases or stroke. β coefficients for the conditions indicated the difference in HRQoL between responders in a disease group and the general population. Subsequently, a base model was estimated adding the EQ-5D-5L dimensions dummies. β coefficients for the conditions indicated the difference in HRQoL between responders in a disease group and the general population, not accounted by the EQ-5D-5L.Differences in β coefficients between the pre-base and the base model indicated the amount of variance in each chronic condition that the 5 dimensions of the EQ-5D were able to explain. Finally, the model was extended including also factors and items, each of which was added individually. If the factor or item took full account of variations in HRQoL for one condition, then the dummy variable for that condition was expected to be insignificant. A reduction in the condition β coefficient represented the responsiveness of the bolt-on to differences in HRQoL for that condition, controlling for the EQ-5D-5L. Non statistically significant β coefficients suggested no impact of the factor or item in predicting differences in HRQoL between patients and the general population. If the addition of a factor or item made one or more of the EQ-5D-5L dimensions not statistically significant, this meant that the factor or item was able to take full account of variations in HRQoL for those dimensions in the general population group. If the factor or item substantially reduced the coefficient of one or more EQ-5D-5L dimensions, this showed a possible interaction between that factor or item and the dimension for which the coefficient was reduced in the general population group.

To enhance clarity of reporting and comparison of results, a maximum of 3 items per factors are presented for the items regressions. Items were selected among those that reported the most prominent results. Full results are presented in Appendices Chapter 5 – Table II.

Analyses of observable variables (item regressions) were conducted in STATA/MP 14©. Analyses of latent variables (factor regressions) were performed in Mplus version 7©.

5.3 Results

Table 5.2 summarizes the percentages of the background characteristics and health status responses, as measured by the 5L version of the EQ-5D, of the survey responders. The following sections summarize the results of the two test employed.

5.3.1 Test 1

Table 5.3 presents the results for the base model using the first test. All EQ-5D-5L dimensions except self care were able to explain variations in HRQoL i.e. health VAS. The number of observations for self care level 5 was small i.e. 8, but the remaining levels reported large number of observations that ranged between 646 (level 2) and 62 (level 4). Removal of usual activities from the regression model resulted in self care level 3 and level 4 becoming statistically significant with coefficients of -3.227 and -5.567, showing a possible interaction between usual activities and self care. β coefficients were larger at increasing levels of severity/ problems for all statistically significant dimensions but for usual activities, where level 4 was associated with a worst decrement than level 5. Mobility reported the smallest β coefficients, which ranged between -3.346 for level 2 and -10.543 for level 5, while anxiety and depression the largest, which ranged between -6.221 for level 2 and -26.102 for level 5.

Figure 5.1 presents the results for the regressions using factors (each added individually). As it can be seen, all factors explained variations in HRQoL over and above the EQ-5D-5L. The size of the β coefficients varied, with coefficients for relationships and satisfaction being approximately double, and of energy/sleep almost triple, than those for the remaining factors. Confidence intervals for energy/sleep, relationships and satisfaction did not overlap, whereas confidence intervals for vision, hearing and speech/cognition did.

All statistically significant dummies for the EQ-5D-5L dimensions in the base model remained statistically significant with the addition of latent factors, with their coefficients registering small or no changes. For example, when energy/sleep was added to the model, EQ-5D anxiety and depression reported a coefficient of -6.221 at level 2 and of -26.178 at level 5 compared to the coefficients of -6.221 at level 2 and -26.102 at level 5 in the base model, while the EQ-5D pain and discomfort reported a coefficient of -4.016 at level 2 and -14.282 at level 5 compared to the coefficients of -4.043 at level 2 and -14.691 at level 5 in the base model.

Figure 5.2 and Table 5.4 report the results for the items regressions (each added individually). As can be seen, items performance differed depending on the factor on which they loaded, with items loading on relationships, satisfaction and energy/sleep registering statistically significant results for most of their levels, and items loading on speech/cognition, vision and hearing being frequently non-significant. β coefficients were generally larger for the items loading on relationships, satisfaction and energy/sleep. For example, while 15D vitality ranged between -9.071 of level 2 and -29.180 of level 5, β coefficients of AQoL hearing were up to more than seven times smaller, and ranged between -1.936 of level 2 and -4.927 of level 4. Similarly, while AQoL close relationships (family and friends) reported a β coefficient of -2.964 for level 2 and -15.310 for level 5, HUI3 cognition β coefficient was -3.022 for level 2 and -7.719 for level 5. AQoL vision registered a large β coefficient for level 4 (-33.079), but this was associated with very large confidence intervals [CI -10.637; -55.521], and is for this reason unreliable.

Systematic differences in items ability to detect variations in HRQoL were seen also between items loading on the same factor. For example, the items measuring energy in the energy/sleep factor i.e. SF-6D vitality, 15D vitality and SF-6D vitality, reported substantially larger coefficients compared to the two items measuring sleep on the same factor i.e. AQoL sleep and 15D sleeping. Similarly, while the two items measuring cognition on the speech/ cognition factor i.e. 15D mental function and HUI3 cognition reported moderate and statistically significant coefficients, all speech items were non statistically significant. The number of observations were similar for the levels of the dummy of these variables e.g. 7 observations for 15D mental function and HUI 3 speech at level 5; 724 observations for HUI3 cognition at level 2 and 641 for 15D speech at level 2.

Finally, some items reported β coefficient decrements that were inconsistent with the increase in the level of severities/problems. For example, 15D mental function reported a larger coefficient for the level 2 of the dummy variable than for the level 3, and AQoL close relationships (family and friend) a larger coefficient for the level 5 than for the level 6. Full results of factor and items coefficients for the first test are presented in Appendices Chapter 5 - table I.

All statistically significant dummies for the EQ-5D-5L in the base model remained statistically significant with the addition of the items. Generally, their coefficients registered small or no changes e.g. when AQoL enjoyment close relationships was added to the model, the EQ-5D usual activities reported a coefficient of -7.382 at level 2 and of -17.190 at level 5 compared to the coefficients of -7.495 at level 2 and -16.764 at level 5 in the base model, while the EQ-5D mobility reported a coefficient of -3.239 at level 2 and -10.222 at level 5 compared to the coefficients of -3.346 at level 2 and -10.543 at level 5 in the base model.

However, coefficients for the EQ-5D anxiety and depression dimension often registered large decrements when items related to satisfaction were added, large to moderate decrements when items related to energy were added and moderate reductions when items related to relationships were added. The greatest reductions were noticed for items related to life satisfaction. For example, when ONS satisfaction with life was added to the model, the EQ-5D anxiety and depression dimension reported a coefficient of -2.346 at level 2 and a coefficient of -13.888 at level 5 compared to the coefficients of -6.221 at level 2 and -26.102 at level 5 in the base model. Similarly, when the SWLS Satisfaction with life was added to the model, the EQ-5D anxiety and depression dimension reported a coefficient of -2.618 at level 2 and a coefficient of -14.317 at level 5. Coefficients for the EQ-5D anxiety and depression dimension with the addition of AQoL energy were -3.261 at level 2 and -16.362 at level 5, while they were -3.633 at level 2 and -20.130 at level 5 when the SF-6D vitality was added. The largest decrement for the EQ-5D anxiety and depression when items related to relationships were added was registered with the AQoL close and intimate relationship (-4.942 at level 2 and -21.717 at level 5).

Coefficients for the EQ-5D usual activities registered moderate decrements when energy items were added. When AQoL energy was added to the model, the EQ-5D usual activities dimension reported a coefficient of -5.045 at level 2 and -10.620 at level 5 compared to the coefficients of -7.495 at level 2 and -16.764 at level 5, while when the SF-6D Vitality was added it reported coefficients of -5.236 at level 2 and -13.525 at level 5.

5.3.2 Test 2

Table 5.5 presents the results of the pre-base model. Table 5.6 presents the base model for the second test. Table 5.7 and Table 5.8 report the change in coefficients associated to the inclusion of factors and selected items. Full results for the second test are available in Appendices Chapter 5 - Table II. In the base model, dummies were statistically significant for all chronic conditions, showing that the EQ-5D-5L only partially captures differences in Health VAS between disease groups and the general population. The smallest coefficients were seen for hearing problems, arthritis and asthma, followed by depression, diabetes and heart diseases. Cancer, COPD and stroke reported the largest coefficients.

None of the factors was able to take full account of differences in Health VAS between patients and the general population, as chronic conditions dummies remained statistically significant and negative for all of them. However, five factors had an impact on one or more of the coefficients of the chronic conditions, reducing their magnitude. More specifically, satisfaction decreased COPD dummy by 0.617 (from -15.570 to -14.953), vision decreased depression, diabetes, COPD and stroke dummies by 0.863 (from -11.123 to -10.260), 0.503 (from -12.565 to -12.035), 0.436 (from -15.570 to -15.134) and 2.623 (from -20.651 to -18.028), and hearing, speech/cognition and energy/sleep decreased stroke dummy by 0.589 (from -20.651 to -20.062), 0.580 (from -20.651 to -20.071) and 1.561 (from -20.651 to -19.090). Decrements for COPD and stroke dummies should be interpreted with care, as they are based on small samples i.e. 23 observations for stroke and 66 for COPD.

All EQ-5D-5L dimensions that were statistically significant in the base model remained statistically significant with the addition of the latent factors, with their coefficients generally reporting small or no changes. For example, addition of energy in the model generated an EQ-5D usual activities coefficient of -6.306 at level 2 and of -15.548 at level 5, which was similar to the coefficients of -6.296 at level 2 and -15.299 at level 5 of the base model. Similarly, addition of relationships in the model generated an EQ-5D anxiety and depression coefficient of -5.074 at level 2 and of -23.282 at level 5, which was in line with the coefficient of -5.074 at level 2 and -23.198 at level 5 of the base model.

Also none of the items was able to take into full account differences in HRQoL between patients and the general population, as chronic conditions dummies remained statistically significant and negative. However, numerous items decreased chronic conditions β coefficients, with some of them having a general impact and other a specific one. For example, AQoL energy produced decrements on all chronic condition dummies that varied between -0.907 for arthritis (from -7.731 to -6.824) and -3.059 for COPD (from -15.570 to -12.551). Similarly, the ONS satisfaction with life reduced all chronic condition dummies with decrements varying between -0.223 for arthritis (from -7.731 to -7.508) and -2.582 for stroke (from -20.651 to -18.069). By contrast, the 15D hearing and HUI3 hearing substantially reduced only hearing problems (decrement of 0.974 passing from -6.890 to -5.916 and decrement of 0.706 passing from -6.890 to -6.184), with the next largest reduction being 4 and 3 times smaller i.e. COPD decreased by 0.226 passing from -15.570 to -15.349 for 15D hearing and by 0.221 passing from -15.570 to -15.354 for HUI3 hearing. Analogously, AQoL close relationships (family and friends) only impacted depression (decrement of 0.667 passing from -11.123 to -10.457), with the second largest coefficient reduction being three times smaller i.e. hearing problems (change of 0.231 passing from -6.890 to -6.659). As previously mentioned, decrements for COPD and stroke should be interpreted with care, as they are based on very small samples i.e. 23 observations for stroke and 66 for COPD.

As for test 1, all the EQ-5D-5L dimensions that were statistically significant in the base model remained statistically significant with the addition of the items. Once again, coefficients for the EQ-5D anxiety and depression dimension often decreased when items loading on satisfaction, relationships and energy were added, with items related to life satisfaction causing the largest switches. For example, when ONS satisfaction with life was added to the model, the EQ-5D anxiety and depression dimension reported a coefficient of -1.366 at level 2 and a coefficient of -11.209 at level 5 compared to the coefficients of -5.074 at level 2 and -23.198 at level 5 in the base model. Similarly, when SWLS satisfaction with life was added to the model, the EQ-5D anxiety and depression reported a coefficient of -1.724 at level 2 and -12.133 at level 5. Coefficients for the EQ-5D anxiety and depression dimension with the addition of AQoL energy were -2.737 at level 2 and -15.392 at level 5, while they were -2.965 at level 2 and -18.563 at level 5 when the SF-6D vitality was added. The largest change in the EQ-5D anxiety and depression coefficients following the addition of items related to relationships was registered with AQoL close relationships (family and friends) (-3.815 for level 2 and -18.967 for level 5).

Items related to energy once again produced decrements in the coefficients for the EQ-5D usual activities. The largest reduction was registered with the addition of AQoL energy (-4.348 for level 2 and -9.887 for level 5) compared to the coefficients of -6.295 at level 2 and -15.299 at level 5 in the base model. With SF-6D Vitality the coefficients of the EQ-5D usual activities were -4.399 at level 2 and -12.443 at level 5.

5.4 Discussion

This chapter reported on a study investigating the potential of using the ability to predict overall HRQoL for selecting between bolt-ons after factor analysis identification. It assessed the usefulness of two tests. The first test aimed at examining the contribution of factors and items in explaining variations of HRQoL not already captured by the EQ-5D-5L; the second test aimed at providing evidence on the areas of health where the factors and items would impact the most if developed/ adapted into bolt-on dimensions.

The first test appeared appropriate for selecting between potential independent factors and independent items. Results for factors and items were concordant in pointing at relationships, energy/sleep and satisfaction factors, and items loading on them, as the mostly relevant bolt-ons. The study also showed systematic differences in items ability to detect differences in HRQoL when they loaded on the same factor. For example, energy items reported consistently larger coefficients compared to sleep items, and cognition items reported consistently larger coefficients than speech items. These results suggest that despite loading on the same factor and being interrelated, energy and sleep and cognition and speech measure partially different concepts. As energy and cognition appeared better in explaining variations of HRQoL than sleep and speech, items related to these concepts should be preferred when adapting them into bolt-on dimensions. These findings agree with those of previous qualitative research (Devlin et al, 2004; Saha et al, 2016) that found "happiness", "emotional health", "cognition", "relationships" and "sensory deprivation" (e.g. vision loss) to be among the most important aspects of health not covered by the EQ-5D. They also provide additional evidence compared to these studies on the relative importance of these aspects as add-on dimensions.

The second test generated results that are sometimes difficult to interpret. While none of the factors and items was able to fully account for differences in HRQoL between patients and the general population i.e. chronic condition dummies remained statistically significant in all the models tested, they were frequently able to explain part of these differences i.e. chronic condition registered reductions after the inclusion of factor and items. Decrements for items generally occurred in chronic conditions that were theoretically related to the aspects of health measured by the item e.g. AQoL hearing reduced hearing problems. Some reductions occurred in chronic conditions not related to the aspects of health measured by the factor e.g. vision reduced COPD. Despite the number of observations for those conditions were generally small and therefore these results should be taken with care, the presence of discrepancies in the results of factors and items regression still pose some doubts on the ability of the second test to discriminate between bolt-ons. Further investigation is needed before using this technique.

This study used a set of strategies that were broadly based on the statistical significance of the factors and items tested, the size and direction of their coefficients and the consistency in HRQoL decrements at increasing levels of severity to discriminate between candidate bolt-ons. The same set of strategies could be also employed to identify dimensions that need bolting off the investigated GPBM. This could be done in isolation, by selecting those dimensions that perform poorly, or comparatively, by comparing the size, direction and interaction of coefficients for the EQ-5D-5L and other measures' dimensions. Using the first approach would suggest, in the case of the current study, to bolt-off the EQ-5D-5L self care dimension, as most of its levels were not statistically significant due to an interaction with usual activities. Using the second approach would suggest, again in the context of the current research, to substitute the EQ-5D-5L self care or the EQ-5D-5L mobility dimensions with a dimension adapted from the items loading on the relationship or the satisfaction factors e.g. AQoL close relationships (including sexual). Choice between these items could be inform by their impact on the remaining dimensions of the EQ-5D, where items causing smaller interactions should be preferred as they tap on aspects less related to those already covered in the EQ-5D. In addition, as some items were seen to report decrements that were inconsistent with the increase in the levels of severity, when choosing an item for adapting it into a bolt-on dimension it would be preferable to use one the levels of which reflect consistent decrements in HRQoL.

Similarly to the latter approach, the size of β coefficients could also be used to set an empirical threshold of bolt-ons relevance. For example, coefficients for the "worst performing item" of the investigated GPBM could be set as a threshold to which coefficients from items that might be adapted into bolt-ons would be compared. If coefficients for the items are at least as large as those of the worst performing dimension in the original measure, then those items would be considered relevant bolt-on additions. This would result in retaining only those items that are at least "as good" as the worst performing item in the GPBMs. However, the usefulness of this approach depends on a number of aspects, among which the method used for the current analysis e.g. linear regression, the measure of HRQoL used i.e. Health VAS and the precision of the model in describing possible interactions between variables.

The approach adopted in this study assumed that bolt-ons reporting larger coefficients i.e. greater ability to explain variations in HRQoL should be preferred to bolt-ons reporting smaller coefficients. However, this is not the only criterion that needs consideration as other fundamental issues exist when selecting bolt-on dimensions, which also include consideration around what is the actual impact of bolt-ons on people’s lives, or what is the effect of bolt-ons on individuals’ preferences for health states taken from the GPBM investigated.

This study has some limitations that need mentioning. First of all, it used the Health VAS as a proxy of HRQoL. The Health VAS reflects self-perceived health and not the value that society attaches to different health states i.e. it reflects a patient view rather than a general population view. This is a limitation, as current practice recommends measuring HRQoL from a general public perspective (e.g. Brazier et al, 2017). Moreover, the EQ VAS has been criticized as a measure of HRQoL as it is not clear what aspects patients consider when responding the question e.g. non-health contextual factors might influence the scores obtained, whether it generates results that are comparable e.g. frequently response shift due to change in perception of the task completed are observed with this measure and whether its endpoints are interpreted consistently between patients (e.g. Feng et al, 2014; McPhail and Haines, 2010). Despite the Health VAS used in this study differs from the EQ VAS, some of the limitations noticed for the EQ VAS might apply also for the Health VAS used in here. Finally, the Health VAS generates values that differ from the TTO (e.g. Brazier et al, 2012; Stolk et al, 2010). If the TTO is chosen as the method of preference elicitation for the EQ-5D to ensure consistency in the valuation methods used for this measure, the choice of testing bolt-ons using the Health VAS may generate sources of mismatch between the selection and valuation phases. Second of all, this study used linear regressions to model the impact of factors and items on HRQoL. Although this technique has been seen in numerous occasions to produce reliable estimates in models with the EQ VAS as dependent variable (e.g. Whynes, 2013; Barton et al, 2009), other models could have been used to account for the bounding of the Health VAS variable e.g. Two limit tobit models. However, this was not considered necessary, as the distribution of the Health VAS was approximately normal. Third of all, possible interactions were noticed among some of the items and the EQ-5D dimensions. Interactions were initially calculated but were not reported as the large number of coefficients generated was of difficult interpretability. Inclusion of interaction terms would have improved the precision of the estimates for those items. Fourth of all, to improve comparability, this study did not use 7 of the items identified as candidate bolt-ons in chapter 4 as they were not related to the factorial structure identified. Moreover, factors and items tested in this study were identified through previous research that did not use disease specific measures. Hence, other factors and items not tested might be equally relevant additions to the EQ-5D-5L if these can be shown to be important to describe general aspects of HRQoL. Fifth of all, the second test performed covered only 9 chronic conditions, but other conditions might have been relevant to better reflect the ability of bolt-ons to detect changes in health. Finally, sample size for the lowest levels of the scale were often small in most of the variables tested. It is important to take estimates generated from these dummies with care.

Despite these limitations, this study presents a useful method to select between alterative factors and items that can be developed/adapted into bolt-on dimensions, as well as providing evidence on the comparative relevance of a set of bolt-ons for the EQ-5D-5L.

Table 5.1 Items wording

	Items	**Wording of the first level of the item**
EQ-5D-5L	EQ-5D-5L mobility	I have no problems in walking about
	EQ-5D-5L self care	I have no problems washing or dressing myself
	EQ-5D-5L usual activities	I have no problems doing my usual activities
	EQ-5D-5L pain/discomfort	I have no pain or discomfort
	EQ-5D-5L anxiety/depression	I am not anxious or depressed
SF-6D	SF-6D vitality	I have a lot of energy all of the time
HUI 3	HUI 3 vision	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glass
	HUI 3 hearing	Able to hear what is said in a group conversation with at least three other people, without a hearing aid
	HUI 3 speech	Able to be understood completely when speaking with strangers or people who know me well
	HUI 3 cognition	Able to remember most things, think clearly and solve day to day problems
AQoL 8D	AQoL energy	[Thinking about how much energy you have to do the things you want to do: I am] Always full of energy
	AQoL close relationships (family and friends)	[Your close relationships (family and friends) are:] Very satisfying

	AQoL communication	[How well can you communicate with others? (e.g., by talking, listening, writing or signing)] I have no trouble speaking to them or understanding what they are saying
	AQoL sleeping	[How often do you have trouble sleeping?] Never
	AQoL enthusiasm	[How enthusiastic do you feel?] Extremely
	AQoL enjoyment close relationships	[How much do you enjoy your close relationships (family and friends)?] Immensely
	AQoL vision	[How is your vision (while using any visual aids you need)?] I have excellent sight
	AQoL hearing	[How is your hearing (while using any hearing aids you need)?] I have excellent hearing
	AQoL close relationships (including sexual)	[Your close and intimate relationships (including any sexual relationships) make you:] Very happy
15D	15D vision	I see normally, i.e. I can read newspapers and TV text without difficulty (with or without glasses)
	15 D hearing	I can hear normally, i.e. normal speech (with or without a hearing aid)
	15 D sleeping	I am able to sleep normally, i.e. I have no problems with sleeping
	15 D speaking	I am able to speak normally, i.e. clearly, audibly and fluently
	15 D mental function	I am able to think clearly and logically, and my memory

		functions well
	15 D vitality	I feel healthy and energetic
ICECAP	ICECAP Love and support	I can have a lot of love, friendship, and support
ONS	ONS satisfaction with life	[Overall, how satisfied are you with your life nowadays?] Completely satisfied
	ONS life is worthwhile	[Overall, to what extent do you feel that the things you do in your life are worthwhile?] Completely worthwhile
	ONS happiness yesterday	[Overall, how happy did you feel yesterday?] Completely happy
PWI	PWI satisfaction with life as a whole	[Thinking about your own life and personal circumstances, how satisfied are you with your life as a whole?] Completely satisfied
	PWI satisfaction standard of living	[How satisfied are you with your standard of living?] Completely satisfied
	PWI satisfaction achievement	[How satisfied are you with what you are achieving in life?] Completely satisfied
	PWI satisfaction personal relationships	[How satisfied are you with your personal relationships?] Completely satisfied
	PWI satisfaction safety	[How satisfied are you with how safe you feel?] Completely satisfied
	PWI satisfaction part of the community	[How satisfied are you with feeling part of your community?] Completely satisfied
	PWI satisfaction future security	[How satisfied are you with your future security?] Completely satisfied

	PWI satisfaction spirituality	[How satisfied are you with your spirituality or religion?] Completely satisfied
SWLS	SWLS life close ideal	[In most ways my life is close to my ideal] How content are you with your life. Strongly agree
	SWLS condition of life are excellent	[The conditions of my life are excellent] How content are you with your life. Strongly agree
	SWLS satisfaction with life	[I am satisfied with my life] How content are you with your life. Strongly agree
	SWLS gotten important things in life	[So far I have gotten the important things I want in life] How content are you with your life. Strongly agree
	SWLS if I could live life over	[If I could live my life over, I would change almost nothing] How content are you with your life. Strongly agree

Table 5.2 – Percentages of background characteristics and health status responses of survey participants

Variable	Category	Percentages
Gender	Male	48%
	Female	52%
Age	18-24	6%
	25-34	12%
	35-44	14%
	45-54	21%
	55-64	25%
	65+	22%
Highest education achieved	High school	31%
	Diploma or Certificate	41%
	University	28%
Chronic condition	None	22%
	Asthma	11%
	Cancer	10%
	COPD	1%
	Depression	11%
	Diabetes	11%
	Hearing problems	10%
	Arthritis	11%
	Heart	12%
	Stroke	1%
EQ-5D mobility	Level 1	67%
	Level 2	19%

	Level 3	10%
	Level 4	3%
	Level 5	1%
EQ-5D self-care	Level 1	88%
	Level 2	8%
	Level 3	3%
	Level 4	1%
	Level 5	0%
EQ-5D usual activities	Level 1	65%
	Level 2	22%
	Level 3	9%
	Level 4	3%
	Level 5	1%
EQ-5D pain discomfort	Level 1	29%
	Level 2	41%
	Level 3	20%
	Level 4	9%
	Level 5	1%
EQ-5D anxiety / Depression	Level 1	50%
	Level 2	29%
	Level 3	14%
	Level 4	5%
	Level 5	2%

Note: each variable has a total number of responders of 8022.

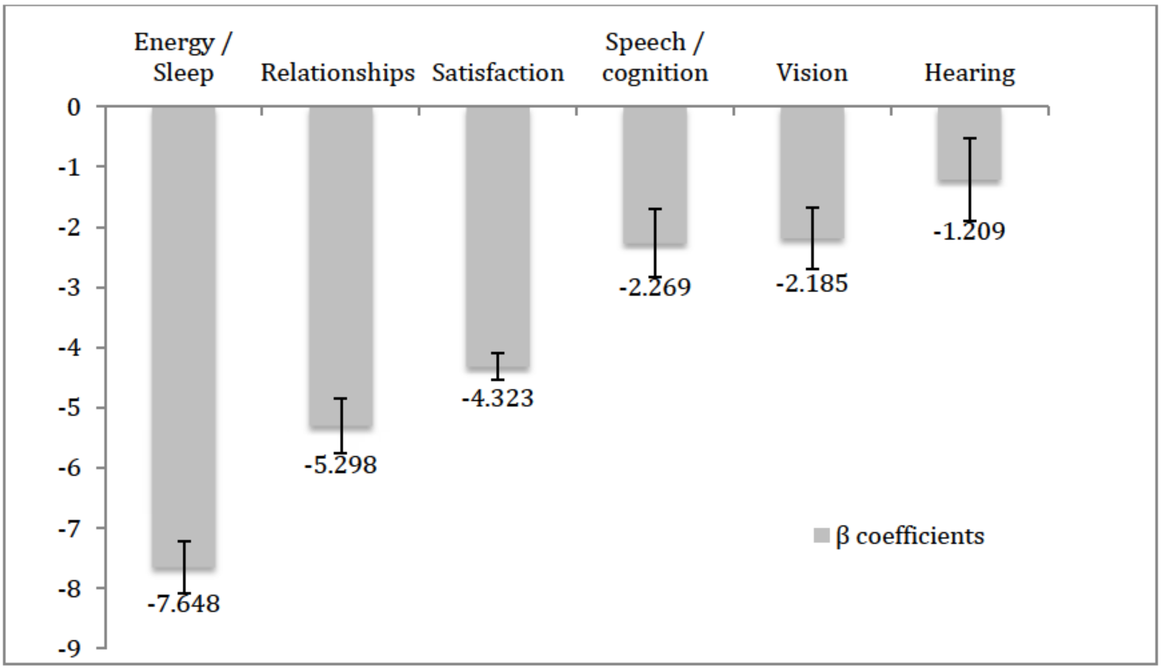
Table 5.3 – β coefficients, statistical significance and standard errors of dummy variables for the base model using the first test (Health VAS dependent variable)

Variables	B coefficients	Standard error
Constant	80.449**	.830
Highschool education	(omitted)	(omitted)
Diploma education	-0.283	.438
University education	0.991*	.482
Age 18-24	(omitted)	(omitted)
Age 25-34	1.105	.912
Age 35-44	0.009	.887
Age 45-54	-0.580	.852
Age 55-64	-0.653	.843
Age >65	1.250	.865
Male	(omitted)	(omitted)
Female	2.729**	.378
EQ5D-5L mobility level 1	(omitted)	(omitted)
EQ5D-5L mobility level 2	-3.346**	.573
EQ5D-5L mobility level 3	-5.788**	.852
EQ5D-5L mobility level 4	-9.479**	1.302
EQ5D-5L mobility level 5	-10.543**	3.249
EQ5D-5L self care level 1	(omitted)	(omitted)
EQ5D-5L self care level 2	-2.202**	.762
EQ5D-5L self care level 3	0.296	1.213
EQ5D-5L self care level 4	-0.941	2.313

	EQ5D-5L self care level 5	1.710	5.924
	EQ5D-5L usual activities level 1	(omitted)	(omitted)
	EQ5D-5L usual activities level 2	-7.495**	.560
	EQ5D-5L usual activities level 3	-12.164**	.900
	EQ5D-5L usual activities level 4	-17.338**	1.428
	EQ5D-5L usual activities level 5	-16.764**	2.584
	EQ5D-5L pain discomfort level 1	(omitted)	(omitted)
	EQ5D-5L pain discomfort level 2	-4.043**	.474
	EQ5D-5L pain discomfort level 3	-7.834**	.646
	EQ5D-5L pain discomfort level 4	-10.341**	.912
	EQ5D-5L pain discomfort level 5	-14.691**	1.776
	EQ5D-5L anxiety depression level 1	(omitted)	(omitted)
	EQ5D-5L anxiety depression level 2	-6.221**	.448
	EQ5D-5L anxiety depression level 3	-12.851**	.603
	EQ5D-5L anxiety depression level 4	-21.522**	.921
	EQ5D-5L anxiety depression level 5	-26.102**	1.378
R^2	0.439		

Note: ** P≤0.05; ** P≤0.01*

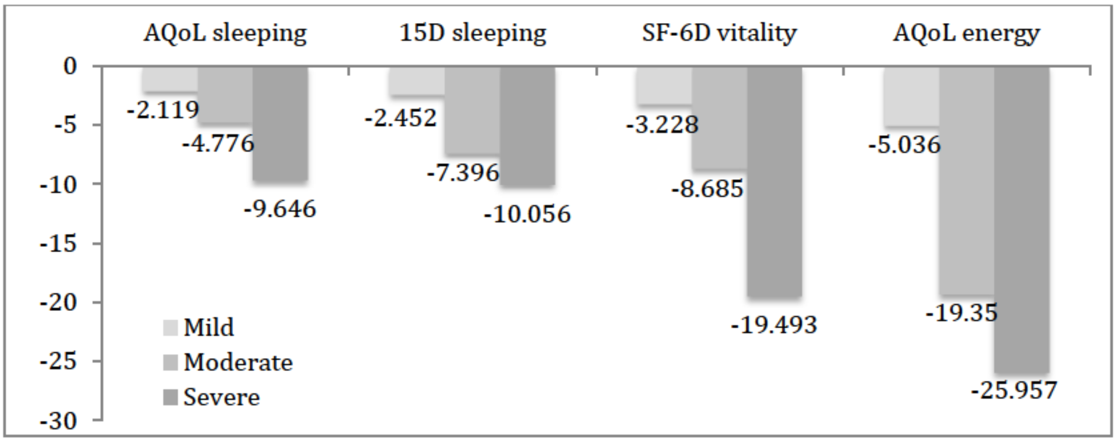
Figure 5.1 β coefficients and 95% confidence intervals of factors using the first test



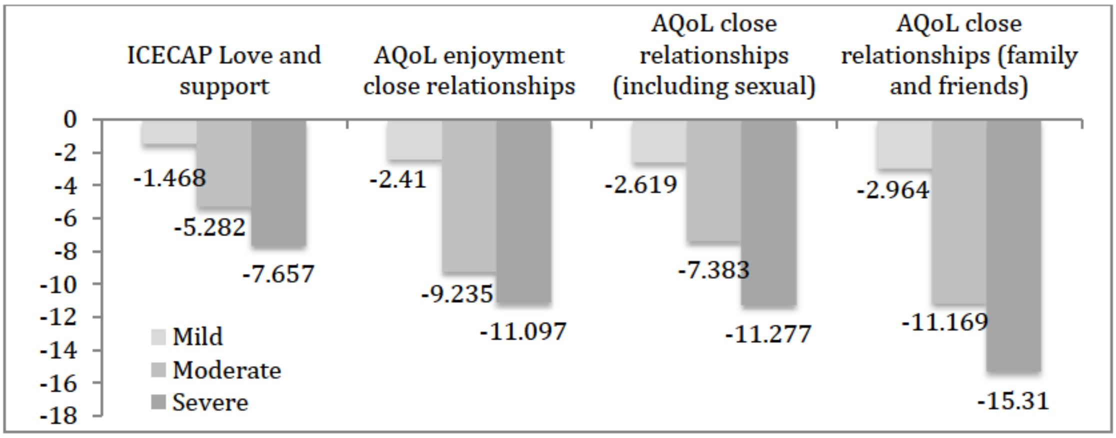
Note: All coefficients are statistically significant at p≤0.01

Figure 5.2 β coefficients for a mild, a moderate and a severe level of selected items using the first test

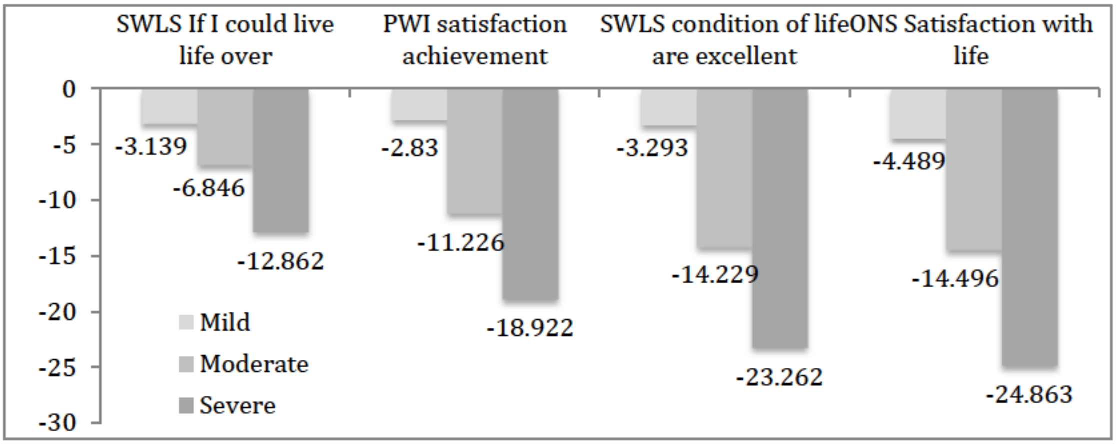
Items loading on energy sleep factor



Items loading on relationships factor

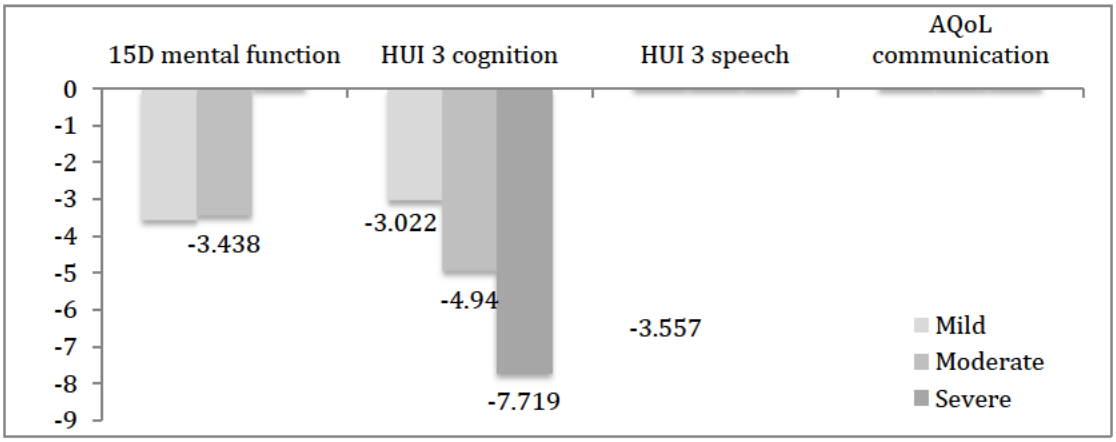


Items loading on satisfaction factor

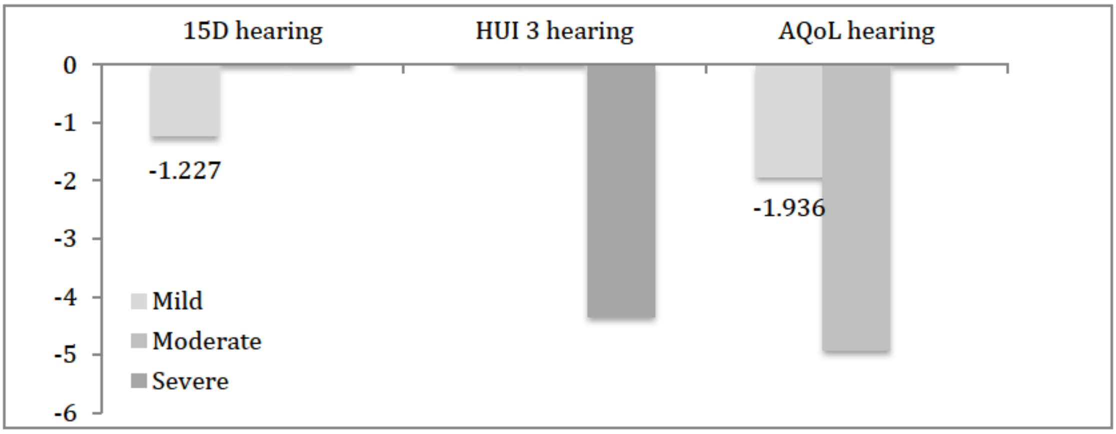


Note: A maximum of four items per factor is presented here. Factors may include additional items; Mild, moderate and severe levels were chosen based on the overall number of levels present in the item; All coefficients are statistically significant at $p \leq 0.05$.

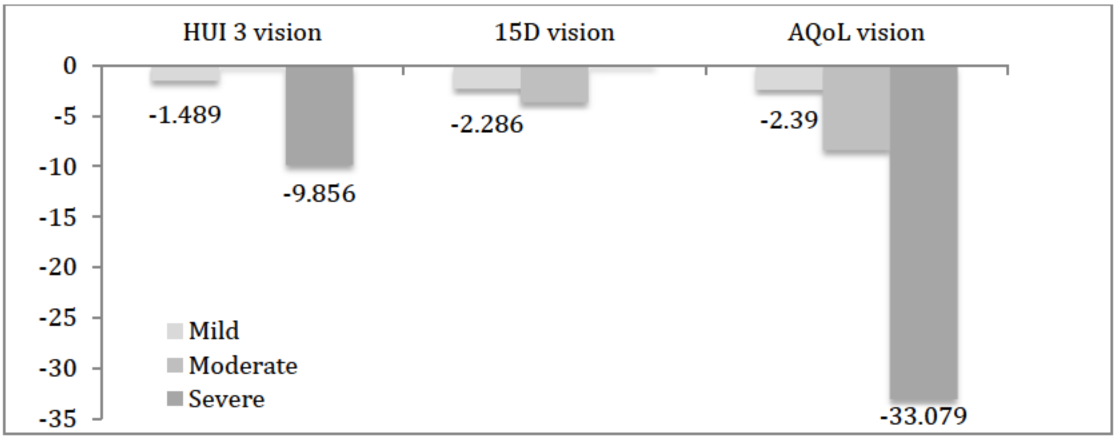
Items loading on speech/ cognition factor



Items loading on hearing factor



Items loading on vision factor



Note: A maximum of four items per factor is presented here. Factors may include additional items;

Mild, moderate and severe levels were chosen based on the overall number of levels present in the item;

All coefficients are statistically significant at $p \leq 0.05$.

Table 5.4 – β coefficients, standard errors and statistical significance of a subset of tested items with EQ-5D-5L for all tested levels

Factor on which item loads	Item tested	VAS Health dependent variable	
		β coefficients	Standard errors
Energy / Sleep	15D vitality 1	*(Omitted)*	*(Omitted)*
	15D vitality 2	*-9.071***	*.456*
	15D vitality 3	*-17.609***	*.665*
	15D vitality 4	*-22.761***	*.862*
	15D vitality 5	*-29.180***	*1.279*
	AQoL sleeping 1	*(Omitted)*	*(Omitted)*
	AQoL sleeping 2	*-2.119***	*.620*
	AQoL sleeping 3	*-4.776***	*.633*
	AQoL sleeping 4	*-7.722***	*.735*
	AQoL sleeping 5	*-9.646***	*.893*
Relationships	AQoL close relationships (including sexual) 1	*(Omitted)*	*(Omitted)*
	AQoL close relationships (including sexual) 2	*-2.619***	*.449*
	AQoL close relationships (including sexual) 3	*-7.399***	*.570*
	AQoL close relationships (including sexual) 4	*-7.383***	*.920*
	AQoL close relationships (including sexual) 5	*-11.277***	*1.365*
	ICECAP Love and support 1	*(Omitted)*	*(Omitted)*
	ICECAP Love and support 2	*-1.468***	*.420*
	ICECAP Love and support 3	*-5.282***	*.604*
	ICECAP Love and	*-7.657***	*1.561*

	support 4		
Satisfaction	SWLS satisfaction with life 1	*(Omitted)*	*(Omitted)*
	SWLS satisfaction with life 2	*-3.511***	*.680*
	SWLS satisfaction with life 3	*-7.887***	*.721*
	SWLS satisfaction with life 4	*-12.342***	*.819*
	SWLS satisfaction with life 5	*-14.587***	*.828*
	SWLS satisfaction with life 6	*-19.591***	*.891*
	SWLS satisfaction with life 7	*-24.868***	*1.145*
	PWI satisfaction standard of living 1	*(Omitted)*	*(Omitted)*
	PWI satisfaction standard of living 2	*-1.010*	*.811*
	PWI satisfaction standard of living 3	*-2.830***	*.751*
	PWI satisfaction standard of living 4	*-6.449***	*.776*
	PWI satisfaction standard of living 5	*-8.040***	*.844*
	PWI satisfaction standard of living 6	*-10.477***	*.870*
	PWI satisfaction standard of living 7	*-9.834***	*.937*
	PWI satisfaction standard of living 8	*-12.078***	*.993*
	PWI satisfaction standard of living 9	*-14.152***	*1.108*
	PWI satisfaction standard of living 10	*-12.871***	*1.402*
	PWI satisfaction standard	*-11.483***	*1.438*

	of living 11		
Vision	AQoL vision 1	*(Omitted)*	*(Omitted)*
	AQoL vision 2	*-2.390***	*.494*
	AQoL vision 3	*-4.070***	*.548*
	AQoL vision 4	*-8.383***	*1.513*
	AQoL vision 5	*-33.079**	*11.450*
	AQoL vision 6	*14.274*	*8.131*
	15 D vision 1	*(Omitted)*	*(Omitted)*
	15 D vision 2	*-2.286***	*.458*
	15 D vision 3	*-3.653***	*1.062*
	15 D vision 4	*-1.448*	*1.595*
	15 D vision 5	*-1.779*	*5.148*
Speech / Cognition	HUI 3 cognition 1	*(Omitted)*	*(Omitted)*
	HUI 3 cognition 2	*-3.022***	*.674*
	HUI 3 cognition 3	*-3.674***	*.560*
	HUI 3 cognition 4	*-4.940***	*1.056*
	HUI 3 cognition 5	*-7.719***	*1.993*
	HUI 3 cognition 6	*5.787*	*4.269*
	HUI 3 speech 1	*(Omitted)*	*(Omitted)*
	HUI 3 speech 2	*-1.445*	*.749*
	HUI 3 speech 3	*.804*	*1.299*
	HUI 3 speech 4	*-2.176*	*2.655*
	HUI 3 speech 5	*-6.090*	*6.177*
Hearing	AQoL hearing 1	*(Omitted)*	*(Omitted)*
	AQoL hearing 2	*-1.936***	*.443*
	AQoL hearing 3	*-2.576***	*.512*
	AQoL hearing 4	*-4.927***	*1.126*
	AQoL hearing 5	*1.584*	*3.571*

	AQoL hearing 6	*8.191*	*5.420*
	15D hearing 1	*(Omitted)*	*(Omitted)*
	15D hearing 2	*-1.227**	*.490*
	15D hearing 3	*-1.360*	*.754*
	15D hearing 4	*-.047*	*2.090*
	15D hearing 5	*1.965*	*4.522*

Note: ** P≤0.05; ** P≤0.01*

Table 5.5 – β coefficients, statistical significance and standard errors of a model with chronic conditions but without the EQ-5D dimensions (Health VAS dependent variable)

Variables	β coefficients	Standard error
Constant	*82.670***	*1.023*
Age 18-24	*(Omitted)*	*(Omitted)*
Age 25-34	*-.140*	*1.071*
Age 35-44	*-2.998**	*1.041*
Age 45-54	*-3.945***	*1.002*
Age 55-64	*-2.215*	*1.002*
Age >65	*2.740**	*1.027*
Highschool education	*(Omitted)*	*(Omitted)*
Diploma education	*1.541**	*.515*
University education	*4.566*	*.565*
Male	*(Omitted)*	*(Omitted)*
Female	*1.266**	*.448*
Cancer	*-24.177***	*.860*
Asthma	*-16.431***	*.814*
COPD	*-35.270***	*2.415*
Depression	*-28.860***	*.799*
Diabetes	*-22.010***	*.808*
Hearing problems	*-11.405***	*.830*
Arthritis	*-20.002***	*.809*
Heart diseases	*-22.847***	*.810*

Stroke	*-37.283***	*4.019*
R^2	*0.483*	

Note: ** P≤0.05; ** P≤0.01*

Table 5.6 – β coefficients, statistical significance and standard errors of the base model using the second test (Health VAS dependent variable)

Variables	β coefficients	Standard error
Constant	*80.449***	*.830*
Highschool education	*(Omitted)*	*(Omitted)*
Diploma education	*0.121*	*.422*
University education	*1.746***	*.465*
Age 18-24	*(Omitted)*	*(Omitted)*
Age 25-34	*1.196*	*.877*
Age 35-44	*0.312*	*.855*
Age 45-54	*0.409*	*.829*
Age 55-64	*-1446*	*.830*
Age >65	*3.551***	*.855*
Male	*(Omitted)*	*(Omitted)*
Female	*2.365***	*.368*
EQ5D-5L mobility level 1	*(Omitted)*	*(Omitted)*
EQ5D-5L mobility level 2	*-2.718***	*.553*
EQ5D-5L mobility level 3	*-5.272***	*.823*
EQ5D-5L mobility level 4	*-9.427***	*1.256*
EQ5D-5L mobility level 5	*-9.955***	*3.126*
EQ5D-5L self care level 1	*(Omitted)*	*(Omitted)*
EQ5D-5L self care level 2	*-2.381***	*.734*
EQ5D-5L self care level 3	*0.043*	*1.168*
EQ5D-5L self care level 4	*-1.955*	*2.225*
EQ5D-5L self care level 5	*1.183*	*5.697*
EQ5D-5L usual activities level 1	*(Omitted)*	*(Omitted)*

	EQ5D-5L usual activities level 2	*-6.296***	*.543*
	EQ5D-5L usual activities level 3	*-10.466***	*.870*
	EQ5D-5L usual activities level 4	*-15.217***	*1.379*
	EQ5D-5L usual activities level 5	*-15.299***	*2.488*
	EQ5D-5L pain discomfort level 1	*(Omitted)*	*(Omitted)*
	EQ5D-5L pain discomfort level 2	*-2.917***	*.462*
	EQ5D-5L pain discomfort level 3	*-6.388***	*.634*
	EQ5D-5L pain discomfort level 4	*-9.293***	*.889*
	EQ5D-5L pain discomfort level 5	*-12.988***	*1.713*
	EQ5D-5L anxiety depression level 1	*(Omitted)*	*(Omitted)*
	EQ5D-5L anxiety depression level 2	*-5.074***	*.435*
	EQ5D-5L anxiety depression level 3	*-10.437***	*.614*
	EQ5D-5L anxiety depression level 4	*-18.642***	*.945*
	EQ5D-5L anxiety depression level 5	*-23.198***	*1.380*
	Cancer	*-14.008***	*.729*
	Asthma	*-9.587***	*.680*
	COPD	*-15.570***	*2.008*
	Depression	*-11.123***	*.753*
	Diabetes	*-12.565***	*.682*
	Hearing problems	*-6.890***	*.688*
	Arthritis	*-7.731***	*.714*
	Heart diseases	*-13.323***	*.686*
	Stroke	*-20.651***	*3.304*
R^2	0.483		

Note: ** P≤0.05; ** P≤0.01*

Table 5.7 Changes in chronic condition β coefficients after factor were included individually (three items per factor)

		Factors – change in coefficients					
Chronic conditions	Base model coefficients	Satisfaction	Relationships	Hearing	Vision	Speech/ Cognition	Energy/ Sleep
Cancer	-14.008	0.045	0.067	-0.069	-0.397	-0.070	0.005
Asthma	-9.587	0.019	0.048	-0.040	-0.206	-0.042	-0.003
COPD	-15.570	**-0.617**	0.294	-0.093	**-0.436**	-0.115	0.638
Depression	-11.123	0.037	0.053	-0.055	**-0.863**	-0.055	-0.022
Diabetes	-12.565	0.062	0.053	-0.074	**-0.503**	-0.071	-0.050
Hearing Problems	-6.890	-0.014	0.031	-0.009	-0.274	-0.014	0.027
Arthritis	-7.731	0.003	0.039	0.026	-0.060	-0.029	0.010
Heart diseases	-13.323	0.047	0.063	-0.069	-0.322	-0.069	-0.030
Stroke	-20.651	1.735	0.010	**-0.589**	**-2.623**	**-0.580**	**-1.561**

Note: Bold indicates reduction in b coefficients ≥0.5. All coefficients were statistically significant at p≤0.05.

Table 5.8 Changes in chronic condition coefficients after items were included individually (three items per factor presented)

		Factors on which items loaded					
		Satisfaction			*Relationships*		
		Items (change in coefficients)					
Chronic condition	Base model coefficients	ONS Life	SWLS Condition life	SWLS Live life over	ICECAP Love, friendship support	AQoL close rel (family friends)	AQoL close rel (sexual)
Cancer	-14.008	**-0.841**	**-0.943**	0.138	0.107	0.174	-0.066
Asthma	-9.587	-0.259	**-0.513**	0.065	0.008	-0.089	0.073
COPD	-15.570	**-0.537**	0.036	0.176	0.179	0.213	0.062
Depression	-11.123	**-0.598**	**-0.883**	**-0.591**	-0.138	**-0.677**	**-0.537**
Diabetes	-12.565	**-0.888**	**-0.832**	-0.170	0.070	-0.118	-0.139
Hearing Problems	-6.890	-0.284	0.045	0.155	-0.074	-0.231	-0.106
Arthritis	-7.731	-0.223	-0.177	-0.165	0.097	-0.02	0.248
Heart	-13.323	**-0.537**	**-0.512**	0.085	0.020	0.027	0.111
Stroke	-20.651	**-2.582**	**-3.519**	**-1.648**	**-1.010**	0.009	-0.473

Note: Bold indicates reduction in b coefficients ≥0.5. All coefficients remained statistically significant at p≤0.05.

Table 5.8 (continued)

		Factors on which items loaded					
		Energy / Sleep			*Vision*		
		Items (change in coefficients)					
Chronic conditions	Base model coefficients	AQoL energy	SF6D vitality	AQoL sleeping	AQoL Vision	HUI3 Vision	15D Vision
Cancer	-14.008	**-1.802**	**-1.048**	-0.434	-0.194	-0.023	-0.052
Asthma	-9.587	**-1.618**	**-1.02**	-0.439	-0.218	-0.043	-0.038
COPD	-15.570	**-3.059**	**-2.187**	-0.451	-0.284	-0.043	0.058
Depression	-11.123	**-2.836**	**-1.975**	**-0.807**	-0.098	-0.101	0.052
Diabetes	-12.565	**-1.914**	**-1.16**	-0.459	-0.323	-0.139	-0.149
Hearing Problems	-6.890	**-1.194**	**-0.561**	-0.296	-0.364	-0.184	-0.208
Arthritis	-7.731	**-0.907**	-0.427	-0.477	-0.187	-0.055	-0.004
Heart	-13.323	**-2.284**	**-1.363**	-0.444	-0.208	-0.063	-0.057
Stroke	-20.651	**-2.401**	-0.265	-0.484	**-0.832**	**-0.751**	-0.477

Note: Bold indicates reduction in b coefficients ≥0.5. All coefficients remained statistically significant at $p \leq 0.05$.

Table 5.8 (continued)

		Factors on which items loaded					
		Hearing			*Speech / Congition*		
		Items (change in coefficients)					
Chronic condition	Base model coefficients	AQoL Hearing	15D Hearing	HUI 3 Hearing	HUI 3 speech	15D mental function	HUI3 cognition
cancer	-14.008	-0.322	-0.112	-0.062	0.001	-0.296	-0.273
asthma	-9.587	-0.165	-0.073	-0.007	0.017	-0.07	-0.101
COPD	-15.570	-0.441	-0.226	-0.221	-0.017	0.08	0.281
Depression	-11.123	-0.153	-0.055	-0.037	0.005	-0.456	-0.440
Diabetes	-12.565	-0.181	-0.089	-0.067	0.016	-0.136	-0.087
Hearing Problems	-6.890	**-1.663**	**-0.974**	**-0.706**	-0.022	-0.327	-0.376
Arthritis	-7.731	-0.182	-0.059	-0.023	0.032	-0.01	0.013
Heart	-13.323	-0.305	-0.108	-0.083	0.01	-0.08	-0.154
Stroke	-20.651	**-0.65**	-0.153	-0.147	-0.139	**-1.137**	**-1.293**

Note: Bold indicates reduction in b coefficients ≥ 0.5. All coefficients remained statistically significant at $p \leq 0.05$.

Chapter 6
Testing the impact of bolt-ons on preferences

6.1 Introduction

Selecting bolt-ons is the product of two criteria. Firstly, does the bolt-on improve the psychometric performance of the EQ-5D? The preferred method to perform this assessment would be testing the EQ-5D psychometric characteristics in multiple conditions, developing the bolt-on dimensions, appending them to the EQ-5D and re-testing the psychometric characteristics of the new extended version of the EQ-5D + bolt-ons in the same populations. This process was not feasible given the constraints of this thesis. As a consequence, the alternative of discriminating between bolt-ons using their ability to predict differences in HRQoL not already accounted by the EQ-5D was used (Chapter 5). While this method is innovative, it does not take into account the second important criteria, namely whether bolt-ons have an impact on preferences for health states. This is a crucial aspect in GPBMs, as their ultimate goal is detecting health decrements that are considered relevant for influencing responders' choices.

This chapter reports the methods and results of a study investigating this second important aspect. The study was structured into two phases. First, factors were developed into bolt-on dimensions and their relevance and face validity was tested using two focus groups. Subsequently, five of the developed bolt-ons were used in a survey. The survey had two objectives: i) to examine whether bolt-ons change preferences for health states described in terms of the EQ-5D; ii) to examine the degree of this change comparatively across bolt-ons to see which are potentially the most important in terms of their likely impact on health state values. The survey used pairwise choices between health states, a method that was chosen as it is simpler and less cognitively demanding (Rowen et al, 2011; Mulhern et al, 2014), as well as cheaper and less time consuming than other techniques commonly used for eliciting preferences e.g. time-trade offs (Brazier et al, 2012). This study was funded through an award received from the EuroQol group in March 2017. The award letter is presented in Appendices Chapter 6 – material I.

6.2 Methods – Development and testing of bolt-ons

6.2.1 Factors versus items

Chapter 4 and 5 referred to factors and items as candidate bolt-ons. However, factors and items are not yet actual bolt-on dimensions, as both of them need to be converted into dimensions for use in a GPBM. Factors do not have descriptors and labels, and for this reason they need their descriptive system to be created from scratch. Items already have descriptors and labels, but these reflect the way in which health is conceptualized in the measure they were taken from (see Chapter 2). For this reason, their descriptive system needs to be adapted to reflect the measure of destination.

Due to the limited resources available for this study, it was judged unfeasible to develop all possible bolt-ons from the factors and items identified in Chapter 4. Hence, an important choice that had to be made was whether to develop new bolt-ons using factors or whether to adapt items into bolt-on dimensions. It was chosen to develop bolt-ons from factors for two main reasons:

i) Factors represent clusters of items. Hence, these are more likely to reflect important constructs that are missing from the parent measure descriptive system.

ii) Descriptors and labels of factors can be developed from scratch, making the construction process easier. By contrast, items wording should be adapted paying attention to avoid alteration that might invalidate the analysis from which they were identified.

Chapter 4 found 6 factors not related to the EQ-5D, and these are: relationships, cognition/speech, energy/sleep, hearing, satisfaction and vision. These factors cover aspects of HRQoL relevant to patients and members of the general public that are not included in the descriptive system of the EQ-5D, but are measured by other GPBMs and SWBMs. For this reason, they represent relevant additions to the EQ-5D descriptive system. Chapter 5 suggested that cognition/speech and energy/sleep are factors covering partially different aspects of health i.e. cognition and speech the first, energy and sleep the second. This was highlighted by the fact that in the regressions presented in Chapter 5 cognition reported larger coefficients that speech and energy reported higher coefficients than sleep, both for factors and for items. Given that these two factors measured two different aspects of health each, it was decided to develop 4 bolt-ons from them. Hence, the 8 bolt-ons developed from the six factors are: relationships, cognition, speech, energy, sleep, hearing, satisfaction and vision. The methods followed for their development and testing are described below.

6.2.2 Development of bolt-ons

Descriptors and labels for each of the investigated factor were developed paying attention to closely resemble the lexical structure of the EQ-5D-5L e.g. Mobility – "I have no problems in walking about", "I have slight problems in walking about", "I have moderate problems in walking about", "I have severe problems in walking about", "I have extreme problems in walking about". This was done to ensure consistency within the measure (see Chapter 3 for more details on why consistency should be ensured).

Wording for the descriptors and labels was assessed by a group of experts composed by John Brazier, Clara Mukuria and Donna Rowen, using the following set of pre-defined criteria: i) consistency across dimensions in terms of conceptualization of health; ii) coherence with the EQ-5D-5L wording; iii) consistency with the construct measured; and iv) ability of the labels to describe plausible health states. In presence of inconsistencies, descriptors and labels were reworded and initial wordings replaced.

Subsequently, alternative possible wordings referred to the investigated constructs were collected from the review of bolt-ons studies presented in Chapter 3. These were compared to the developed wordings using the same criteria mentioned above. If it was not possible to establish a superior alternative between options, multiple wordings were carried onto the next phase.

6.2.3 Testing of bolt-ons relevance and face validity

A bottom up approach i.e. involving patients and members of the general public was adopted to test bolt-on relevance and face validity (see Chapter 3 for a definition of relevance and face validity). The reason is two fold. On the one hand, bottom up techniques have been seen to produce questionnaires with a more appropriate language for the population of interest, and therefore to increase questionnaire's face validity (Brod et al, 2009). On the other hand, techniques that involve patients and laypeople in the development of HRQoL measures are receiving increasingly more interest and recognition (Patrick et al, 2011) as a result of the general move of modern health systems towards patients' centred healthcare (Institute of Medicine, 2001; Arah et al, 2006).

A choice had to be made between collecting data through interviews or focus groups. If on the one side interviews have the advantage of offering more in depth information, on the other focus groups allow participants to feed off each other's ideas and tend to stimulate discussion (Stevens and Palfreyman, 2012). As discussion within the group was considered more relevant than individuals' opinions for identifying ambiguities in the wording of descriptors and labels of the dimensions, the focus group technique was chosen. This technique has also the additional advantage of being cheaper and less time consuming compared to individual interviews, which responded to the need of adopting a pragmatic approach for this phase of the thesis.

Two focus groups were conducted, one with members of the general public and one with patients. The choice of using these two populations is motivated by the fact that GPBMs have been developed for use in both these populations. The School of Health and Related Research ethics committee reviewed and approved the first focus group in date 10/10/2016, and the second focus group in date 26/01/2017. Ethics approval letters are available in the Appendices Chapter 6 – material II and Appendices Chapter 6 – material III.

The target size for both focus groups was of 5 to 8 participants. These were recruited through the students and staff volunteer list of the School of Health and Related Research at the University of Sheffield using a convenience sampling technique. Individuals were eligible to participate in the first focus group i.e. general population if they: i) were more than eighteen years old; ii) declared to be fluent in English. Individuals were eligible to participate in the second focus group i.e. patients if they: i) were more than eighteen years old; ii) declared to be fluent in English; iii) declared to be affected by a chronic health condition.

Participants were firstly administered the EQ-5D-5L questionnaire and were asked to complete it. They were then asked to review their answers, describe what they had thought about while completing the questionnaire and comment on whether they thought the EQ-5D was missing domains relevant for their HRQoL. This section provided the opportunity for participants to familiarise with the measure.

Subsequently, participants were administered the bolt-ons grouped by theme. The order with which bolt-ons were presented followed the topics emerged during the first part of the focus group e.g. if participants mentioned the absence of sensory dimensions, sensory theme bolt-ons were presented first, followed by the other themes. This phase collected ideas on bolt-ons relevance and face validity i.e. clarity, ease of understanding and responding to and acceptability in different populations. Probes from the topic guide were used to aid discussion on the topics investigated.

The second focus group included a third phase in which bolt-on variants developed based on the first focus group suggestions were administered. The additional variants were presented as a third phase to ensure comparability in the results of the first 2 phases between the 2 focus groups.

The focus groups were audio-recorded and transcribed, with the data being anonymised using pseudonyms. For the analysis, participants' statements were initially grouped into themes. Subsequently, thematic content analysis (e.g. Grbich, 1999) was used to investigate further the issues emerged in the discussion. Results of both focus groups were used to finalize the wording for the descriptors and labels of the 8 bolt-on dimensions.

The invitation letters, information sheets, background questionnaires, topic guide and the EQ-5D questionnaire for the two focus groups are presented in Appendices Chapter 6 – materials IV – XI.

6.3 Methods – Pairwise choice survey

A pairwise comparison survey was developed and administered to a representative sample of the UK general population to test the impact of the developed bolt-ons. The next section describes some of the key aspects of the experiment.

6.3.1 Selection of bolt-ons

Only 5 of the 8 bolt-on variants developed through the focus groups were used in the survey due to resource limitations. Bolt-ons examined were hearing, sleep, cognition, energy and relationships. Relationships, energy and hearing were selected as in the regressions presented in Chapter 5 they reported large, moderate and small coefficients respectively. Using these bolt-ons allows comparison of results between the regression study and the current study. Cognition and sleep were selected as the first was seen to have a large impact on preferences for health states in Krabbe et al (1999), while the second one to have no impact on preferences for health states in Yang et al (2014). Using these bolt-ons allows comparisons of the current study results with those of previous bolt-on research.

6.3.2 Selection of EQ-5D-5L health states pairs and bolt-on levels

Three pairs of health states were chosen based on the EQ-5D-5L value set for England study (Mulhern et al, 2014). To enhance interpretability of results, pairs were selected among those where responders' preferences were in approximately equal proportions i.e. 50% of responders preferred health state A and 50% health state B. The selected pairs of health states are: health state pair 1 (State A 11122 versus State B 23111); health state pair 2 (State A 52211 versus State B 11325); health state pair 3 (State A 33142 versus State B 34333).

There are 25 possible combinations of bolt-on levels for each bolt-on for each pairwise choice i.e. 1 vs. 1; 1 vs. 2; 1 vs. 3 etc. Due to resource limitations, it was considered unfeasible to test all possible combinations. Hence, three levels per bolt-on were chosen for this study out of the possible five, and these were levels one, three and five. The first level was included since it allows an assessment of whether the simple presence of a bolt-on dimension changes preferences for the pairs of health states presented. The second and third combinations were selected as they allow investigating the importance of different bolt-ons in terms of their impact on health states preferences for moderate and severe levels.

6.3.3 Design

Responders were asked to express their preference between pairs of health profiles. All health profiles included the five areas of health described by the EQ-5D-5L, namely mobility, self-care, usual activities, pain /discomfort, anxiety/depression. Some of the health profiles included an additional dimension of health from the selected bolt-ons: hearing, sleep, relationships, energy and cognition.

The study presented responders with selected pairwise choices of health states. Pairwise choices were chosen as the preferred method as they are simpler and less cognitively demanding than other methods (Rowen et al, 2011; Mulhern et al, 2014), which mean they can be undertaken without an interviewer online. Online administration was preferred as it is cheaper and less time consuming than other techniques commonly used for eliciting preferences e.g. time trade-off (Brazier et al, 2012).

A block design was employed for the survey. Each block included eight pairwise questions. This was considered a feasible number of tasks per participant based on previous research (e.g. Devlin et al, 2016). Three of the pairwise questions did not include bolt-ons, while the remaining 45 pairwise questions included one bolt-on option, making a total of 48 pairwise questions. Each pairwise comparison asked respondents to select the profile they preferred (An example of the pairwise question for a bolt-on at level 3 is presented in Figure 6.1). Participants allocated to the blocks 1, 2 and 3 completed one task comparing pairs of EQ-5D-5L states without bolt-ons in each block and 7 tasks comparing pairs of EQ-5D-5L states with bolt-ons. Participants allocated to the blocks 4, 5 and 6 completed 8 tasks comparing pairs of EQ-5D-5L states with bolt-ons. Bolt-ons at severity 3 and 5 were always added on health states 11122, 52211 and 33152. No indifference option was provided, as done in previous research (Mulhern et al, 2014). To avoid focusing effects, each block included a mix of health state pairs and bolt-ons. The survey presented two levels of randomization. Firstly, participants were randomized in one of the 6 blocks. Subsequently, a randomization of the side in terms of which options participants saw as option A and option B was performed to avoid any position bias.

6.3.4 Sample

It was estimated that in order to detect a 10% difference between responses with and without a bolt-on dimension, using a two-sided test with power of 0.8 and significance level of 1%, 340 responders per pairwise comparison was required (Bansback et al, 2014). Hence, the target sample for the survey was of 1020 participants. The sample aimed at presenting a full distribution of age categories (18+) and an equal proportion of participants between genders.

Participants were recruited using an existing UK online panel administered by Research Now®, a market research company. This included individuals that have previously signed up to answer surveys in return for points that can be exchange for goods. Each responder used a web link to access the survey and was for this reason able to self-complete it at his/her own convenience. The survey had four components presented in the following order: (1) Background and socio-demographic questions; (2) Self-reported health assessed through the EQ-5D-5L + bolt-ons; (3) Familiarisation session; (4) Survey. Appendices Chapter 6 - materials XII - XV present the invitation e-mail, the information sheet, the consent form and the survey.

6.3.5 Analyses

The background characteristics of the participants allocated to the different blocks were compared. Chi square tests and Fisher's exact test were used to assess whether there were statistically significant differences in age, gender, social and economic status across the 6 blocks. Then, frequencies and percentages of responders' preferences between the two profiles (i.e. A and B) for each of the tasks were compared, after adjusting for randomization of the side.

To assess whether bolt-ons changed preferences for EQ-5D-5L health states, statistical testing of differences in proportions of responses between the three pairwise choices without bolt-ons and the corresponding pairwise choices with the bolt-ons was performed using Z tests. As the EQ-5D-5L pairs were chosen based on an approximately equal chance of selection, it was expected that the addition of bolt-ons at severity levels 3 and 5 would result in a lower proportion choosing states with these particular levels of severity. In order to discriminate and inform the selection between bolt-ons, differences in the proportions of preferences between bolt-ons at the same level for the same health state pairs were tested using Z tests e.g. the proportion of responders choosing option A for hearing at level 5 for pairwise option 11122 - 23111 was compared to the proportion of responders choosing option A for relationships at level 5 for the same pairwise option. A bolt-on was considered as having a larger impact on preferences for the EQ-5D-5L in comparison to another if the two ways Z test showed a smaller number of responders choosing option A for that bolt-on, and this difference was statistically significant.

6.4 Results – Bolt-ons development and testing

6.4.1 Bolt-ons development

For 5 of the 8 factors, life satisfaction, relationships, hearing, vision and sleep, a single variant of bolt-ons labels was developed. For vision, this reproduced the wording chosen by Yang et al (2015) modifying it to fit the 5L version of the EQ-5D. For speech, 2 bolt-on variants were produced. The remaining 2 bolt-ons, energy and cognition, presented 3 variants of the bolt-on. For energy, one of these variants reproduced the wording chosen by Yang et al (2015) modifying it to fit the 5L version of the EQ-5D. Bolt-on variants are reported in table 6.1.

6.4.2 Bolt-ons relevance and face validity

The general population focus group recruited 5 participants (AC, LB, MH, PN, SR) and the patients' focus group 6 (DM, HS, LC, LL, QB, TI). Frequencies of their background characteristics are presented in table 6.2. All responders in the general population focus group were female, while in the patients' focus group half of them were male. Participants in the general population focus group were aged between 18 and 34, while in the patients focus group 4 were older than 35. Two general population focus group responders had had an experience with illness among their family and friends. Patients were affected by different diseases, and more specifically 2 had chronic obstructive pulmonary disease, 2 type 1 diabetes, 1 chronic fatigue syndrome and 1 endometriosis. In the general population focus group 1 participant declared to be religious, while in the patients focus group 4. Only 1 responder had had a previous experience with the EQ-5D, and this was in the patients' focus group.

Initial discussion on the EQ-5D

Participants in both focus groups identified a number of aspects of health not already captured by the EQ-5D. Two participants (LB and LC), one in the general population and one in the patients' focus group, argued that an important aspect not covered by the EQ-5D was sleep. Two other participants (AC and QB), one in the general population and one in the patients focus groups, suggested that relationships was not covered in the questionnaire. One participant in the patients' focus group considered concentration or focus an aspect of health that would be relevant to add. One last participant in the general population focus group claimed that the mental health section needed expansion.

Bolt-ons relevance

Only two of the bolt-on dimensions developed from the identified factors, relationships and sleep, where considered relevant aspects to describe HRQoL by responders in both focus groups. Among the remaining bolt-ons, life satisfaction was considered a very relevant aspect of HRQoL by the general population responders, but not by patients. By contrast, patients considered cognition, vision, hearing, speech and energy relevant aspects of HRQoL, while general population responders did not consider these dimensions important.

Bolt-ons face validity

Participants in both focus group found hearing, sleep and the two variants of the speech bolt-on easy to understand, easy to respond to and consistent with the remaining dimensions of the EQ-5D. It was noted that the two variants of the speech bolt-on i.e. *"I have no problems with my speech"* and *"I have no problems speaking"* described different problems, with the first one being more related to the physical ability of pronouncing words, and the second one being more related to the confidence in expressing yourself. Patients preferred the first option while general population responders the second option.

Participants of both focus groups criticised the wording of the descriptor and labels of the relationships bolt-on as these were considered too vague. For example, PN reasoned:

"Does [relationships] mean with your family? With your friends? […]. I can have really good relationships with my partner but […] I just do not want to keep in touch with my family or his family…".

Similarly, LC stated:

"[*as the question is posed] you could think: well I am going through a divorce so I am having loads of problems with that […]. It is your quality of life but it has nothing to do with your health*"

The new wording for the labels of the relationship bolt-on developed following the comments of the general population focus group (presented in Table 6.3) was considered even more vague and difficult to respond to compared to the first version for responders in the patients' focus group.

Participants in the two focus groups disagreed on whether the remaining bolt-ons were easy to understand and respond to. For example, the general population responders found the descriptor and the label at the 5th level of the vision bolt-on i.e. *"I have extreme problems seeing"* unclear. As regards the descriptor, this was criticised as it did not cover the possibility of using glasses. The label was considered inconsistent with the remaining dimensions of the EQ-5D. By contrast, patients found no problem with the original version of the vision bolt-on, but found the new version developed following the general population focus group comments (Table 6.3) inappropriate to describe vision for those not wearing glasses.

Similarly, general population responders criticised the labels wording for the 3rd and 4th level of the life satisfaction bolt-on. This point is well explained by MH:

"I would probably say…I am quite satisfied with my life, and instead of moderately I would just say I am satisfied with my life… cause there is too many adverbs there".

While the original version of the bolt-on appeared easy to understand and to respond for the patients focus group, the new version developed based on the general population comments (Table 6.3) was considered inconsistent with the EQ-5D and therefore difficult to understand. For example, when presented with the new version DM commented:

"*In the original one, the first one, I ticked moderately satisfied. When I came to the middle one of this one [second variant], I am satisfied, to me that was not the same thing. Saying I am satisfied means everything is fine. That would be almost the top. So I did not tick that one, I ticked the one down. I prefer to put moderately satisfied".*

Finally, both cognition and energy were criticised by the general population responders in all their variants, as these two aspects were not considered applicable to them. By contrast, patients argued that in both cases all three variants were consistent with the structure of the remaining dimensions of the EQ-5D and were easy to understand and respond to. They expressed a preference for the variant using the labels *"I have no problems with remembering things"* for cognition and *"I have no problems with my energy levels"* for energy.

Following these results, 2 bolt-ons were modified. The phrase *(while using any visual aid you might need)* was added to the descriptor of vision and its label for the 5^{th} level was changed from *"I have extreme problems seeing"* to *"I am unable to see"*. These modifications responded to the problems raised in the general population focus group while simultaneously accommodating for the interpretability issue rose when the new bolt-on version was presented in the patients' focus group. Relationships was modified specifying in the labels of each of its levels "social relationships". This change was made to solve the fact that the bolt-on was found too vague in both focus groups.

The remaining bolt-ons were not modified. For hearing, sleep and speech, this choice was made as both focus groups did not report problems with their descriptors and labels. For energy and cognition no changes were made as the patients' focus group showed these were easy to understand and to respond to. Satisfaction was not modified as patients expressed a preference for the original wording, considering the new bolt-on developed after the general population focus group difficult to understand and to respond to.

A choice had to be made for those bolt-ons for which multiple variants were presented. For speech, the first option i.e. *"I have no problems with my speech"* was carried forward as characteristics of the individuals such as confidence were not judged as an important element to describe HRQoL. Following the suggestions of the patients' focus group, the first option was chosen for the cognition i.e. *"I have no problems with remembering things"* and the energy bolt-ons i.e. "*I have no problems with my energy levels"*. Table 6.4 presents the 8 bolt-on dimensions finalized after the focus groups.

6.5 Results – The pairwise choice survey

The survey was undertaken in May 2017. 1581 individuals entered the survey. 342 were excluded as they did not select all options in the consent form and so could not be included for ethical reasons, 169 as they did not complete the survey, 5 as they "speeded" through the survey (threshold for speeding is calculated as the median completion time divided by 3) and 25 as their quota was already full. The final analysis set comprised 1040 participants. The mean time taken to completion was 9.19 minutes (range 2.12 - 245.33 minutes), and the median time 7.05 minutes. Participants in block 2 took the shortest mean time (7.23 minutes), while participants in block 4 the longest mean time (10.52 minutes). The background characteristics of the sample are presented in Table 6.5, and their self-reported health status in Table 6.6. No statistically significant differences were found between participants allocated to the six blocks in terms of gender, marital status, profession and highest education achieved. Differences in age were seen between block 1 and 4 ($p<0.001$), with block 1 appearing more normally distributed with a strong central tendency at age 45-54, and block 4 presenting a relatively uniform number of responders in each age category. Self-reported health status was generally similar across blocks except for responders in block 2 who reported more problems in self-care than responders in block 6, and this difference was statistically significant ($p=0.00$).

Table 6.7 presents the frequencies and percentages of responders' choosing option A and option B for the three pairs of health states after adjusting for randomisation of the side, with and without bolt-ons. Only one of the three selected EQ-5D-5L pairs (Health state pair 2) had a close to 50:50 selection of either health state in the pair when no bolt-ons were included. Choice A was the preferred option for the other two pairs with no bolt-on.

Additions of bolt-ons at level 1 did not appear to affect responders' choices, with variations between questions with and without bolt-ons ranging between ±10 percentage points and being not statistically significant. For example, in health state pair 2 51.7% of responders chose option A without bolt-ons, and 52% option A with energy at level 1 ($p=0.96$). Similarly, in health state pair 3 70.6% of responders chose option A without bolt-ons, and the same number of responders expressed a preference for option A when hearing was added at level 1 ($p=0.99$). Only in two questions differences were larger than ± 10%, and these were relationships and sleeping in health state pair 3, with differences being statistically significant only for relationships ($p=0.01$).

The addition of a bolt-on at level 3 generally decreased the number of responders choosing option A compared to the same question without bolt-ons. These reductions ranged between 29.4% (cognition) and 15.4% (energy) in health state pair 1, and 34.3% (cognition) and 18.4% (relationships) in health state pair 3. All differences were statistically significant at $p<0.05$. The reduction in proportions in health state pair 2 were smaller, ranging between 3.4% (hearing) and 13.0% (relationships). These were statistically significant for cognition and relationships, but borderline for sleeping ($p=0.06$) and not at all for hearing ($p=0.52$). Energy registered a higher number of responders choosing option A for the pairwise choice with the bolt-on at level 3 than the corresponding pairwise choice without bolt-on in health state pair 2, but this difference was not statistically significant.

The impact of adding a bolt-on at level 3 varied by dimension. As it can be seen in Table 6.8, cognition produced the largest switch in preferences in two of the three health state pairs (1 and 3), and ranked 2nd in health state pair 2. By contrast, energy at level 3 had the least impact on preferences in health state pairs 1 and 2, and the second least impact in health state pair 3. Sleeping and hearing reported similar impacts, and ranked 2nd and 3rd in health state pair 1 and 3, and 4th and 5th in health state pair 2. Relationships results were inconsistent, registering the smallest impact on preferences in health state pair 2, the second smallest impact in health state pair 1 and the largest impact in health state pair 3. Differences in the proportion of responders choosing option A were statistically significant between cognition, and energy and relationships in health state pair 1, and between cognition and energy, relationships and sleep in health state pair 3. The proportion of responders choosing energy was statistically significantly larger than the proportion of responders choosing sleep and hearing in health state pair 1 and than sleep, cognition and relationships in health state pair 2. This seems to suggest that cognition would register the largest coefficient for its level 3 in a valuation study, while energy the smallest.

As expected, the addition of a bolt-on at level 5 always decreased the number of responders choosing option A more than at level 3. Differences from the question without bolt-ons ranged between 53.2% (hearing) and 31.1% (sleeping) for health state pair 1, 31.1% (hearing) and 17.2% (cognition) for health state pair 2, and 50.5% (hearing) and 25.4% (sleeping) for health state pair 3. All differences were statistically significant at p≤0.05. Differences from the question with bolt-on at level 3 ranged between 28.4% (hearing) and 6.1% (sleeping) in health state pair 1, between 27.7% (hearing) and 6.3% (cognition) in health state pair 2, and between 24.6% (relationships) and 5.4% (sleeping) in health state pair 3.

As with level 3, comparison of responders' choices for bolt-ons at level 5 revealed that these differed depending on the bolt-on. As it can be seen in Table 6.9, hearing consistently reported the largest impact on preferences in all three health state pairs. By contrast, sleeping appeared as the most favourable option in two of the three health state pairs (1 and 3), and registered a high percentage of responders choosing health state A also in health state pair 2. Cognition, energy and relationships registered a similar impact in all three pairs, with cognition having a marginally larger impact in health state pair 1 and 3. Differences in response preference were statistically significant between hearing, and energy, sleep and relationships in health state pair 1, between hearing, and energy, sleep, cognition and relationships in health state pair 2, and between hearing, and energy and sleep in health state pair 3. Differences between sleeping, and energy and cognition, were statistically significant in health state pair 1 and 3 at $p<0.05$. Responders' choices for energy, cognition and relationships were substantially similar, and differences between them were not statistically significant. This seems to suggest that hearing would register the largest coefficient for level 5 in a valuation study, sleeping the smallest and energy, cognition and relationships a similar coefficient.

6.6 Discussion

This chapter presented the results of a study investigating the potential of using pairwise comparisons to determine whether bolt-on dimensions previously identified through factor analysis change preferences for EQ-5D-5L health states. The aim was to test the use of simple low cost pairwise comparisons as a method for selecting bolt-on dimensions.

The study showed that each of the individual bolt-ons had a significant impact on preferences for the EQ-5D-5L. The extent of this impact varied according to the bolt-ons and their severity level, as well as the health states to which they were added. Additions of bolt-ons at level 1 generally resulted in differences of ± 10 percentage points compared to the same pairwise choice without bolt-on, with these differences not being statistically significant, as expected. Additions of bolt-ons at level 3 generally produced a reduction in the percentage of individuals choosing the health state to which the moderate level was added. Additions of bolt-ons at level 5 generally resulted in a smaller number of responders selecting that health state compared to both the same health state without bolt-on and the same health state with a bolt-on at level 3. The dimensions that had the largest impact were hearing and cognition, while sleep and energy had less impact.

These findings agree with those of previous research in that they show that hearing and cognition make a significant impact on the judgments people place on the EQ-5D health states (Krabbe et al, 1999; Yang et al, 2015). These findings also show that sleep has less impact on preferences for EQ-5D health states, which is consistent with the findings of Yang et al (2014).

This study found that at severity level 5 hearing had the largest impact on responders' evaluations, followed by cognition, relationships and energy with similar impacts, and sleeping with the smallest impact. By contrast, at severity level 3 cognition caused the largest switch in preferences, followed by hearing, relationships and sleep, with energy registering the smallest switch. This suggests that the relative weight responders place on different health problems is not constant across levels of severity between bolt-ons. This is relevant for selecting bolt-on dimensions, as it highlights the need for a judgment on what decision rule needs to be followed. One possibility might be choosing the bolt-ons that causes the greatest switch in preferences compared to same health state without bolt-on i.e. choosing based on the worst severity level. Alternatively, bolt-ons might be selected based on the mean change in preference they cause between different levels. Either way, other considerations remain fundamental for the final selection, such as what other dimensions are already present in the descriptive system of the examined measure, and the reason for adding the bolt-on dimensions.

This study selected three pairs of health states where responders' preferences were in approximately equal proportions based on the EQ-5D-5L value set for England study. However, two of the three pairwise choices, health state pair 1 and health state pair 3, generated substantially different results from the ones of the value set for England study. This might be due to differences between the two samples in terms of background characteristics, as well as differences in responders' preferences. Either way, this study results were not affect by these differences, as it was still possible to detect statistically significant variations in terms of preferences switch for the pairwise choices.

This study has a number of limitations that need mentioning. First of all, it was chosen to select bolt-ons from factors and not from items. Despite this choice due to the time and resource constraints of this thesis, it is reasonable to assume that this might have excluded items the content of which might be equally important for general public or patients preferences. Second of all, bolt-ons descriptors and labels were refined using two focus groups. However, the focus groups sample was small and patients and general population responders often disagreed on the relevance and face validity of the bolt-on options presented. As an extensive investigation of the wording of descriptors and labels was not conducted, the bolt-ons tested in the survey cannot be considered definitive. More systematic qualitative studies using larger samples are recommended before these can be appended to the EQ-5D. Third of all, data for the survey were collected online. On the one hand, this mode of administration does not allow to fully understanding whether participants engaged in the choice task and what was their level of concentration beyond the time taken to complete the survey. Some methods have been proposed to assess the level of engagement of participants e.g. eye tracking (e.g. Krucien et al, 2014), but these were not used in the current study due to resource limitations. Nevertheless, the generally easily interpretable and consistent results suggest that responders had a reasonable level of engagement. On the other hand, this method of administration has been seen to produce responses with greater variations than face-to-face administration for a number of elicitation techniques (e.g. Norman et al, 2010; Robinson et al, 2008). Therefore, use of face-to-face administration might have improved the results obtained. Fourth of all, this study used pairwise choices to elicit preferences. This decision relies on the assumption that a large switch in preferences in the pairwise choice tasks would result in a large decrement in the coefficient associated to that level of the bolt-on in a valuation study. However, pairwise choices generate different utility values from the TTO (e.g. Brazier et al, 2012; Stolk et al, 2010). As NICE recommends the use of the EQ-5D with health state values obtained from the TTO (e.g. Devlin et al, 2011), different results from those obtained in this study might be observed in an actual valuation study. Fifth of all, this study selected only three pairs of health states. Previous research has shown that preferences for bolt-ons might vary depending on the severity of the health states to which they are added (Yang et al, 2015) and for this reason other pairs might have generated different results. Sixth of all, only a subset of the levels of the bolt-ons was tested. As it was noted that the relative weight responders place on different health problems is not constant but rather depends on the severity level of the additional dimension, further testing of level 2 and 4 for the same bolt-ons is required to confirm these findings. Despite these limitations, this study provides important evidence in that it proposes a useful and easy method for selecting bolt-on dimensions. Further research is recommended on testing other bolt-ons and all levels of the included ones.

Table 6.1 - Candidate bolt-on variants for general population and patients focus groups

LIFE SATISFACTION	
I am extremely satisfied with my life	☐
I am very satisfied with my life	☐
I am moderately satisfied with my life	☐
I am slightly satisfied with my life	☐
I am not satisfied with my life	☐

RELATIONSHIPS	
I have no problems with relationships	☐
I have slight problems with relationships	☐
I have moderate problems with relationships	☐
I have severe problems with relationships	☐
I am unable to have relationships	☐

HEARING	
I have no problems hearing	☐
I have slight problems hearing	☐
I have moderate problems hearing	☐
I have severe problems hearing	☐
I am unable to hear	☐

SPEECH (option 1)

I have no problems with my speech ☐

I have slight problems with my speech ☐

I have moderate problems with my speech ☐

I have severe problems with my speech ☐

I am unable to speak ☐

SPEECH (option 2)

I have no problems speaking ☐

I have slight problems speaking ☐

I have moderate problems speaking ☐

I have severe problems speaking ☐

I am unable to speak ☐

VISION

I have no problems seeing ☐

I have slight problems seeing ☐

I have moderate problems seeing ☐

I have severe problems seeing ☐

I have extreme problems seeing ☐

COGNITION (option 1)

I have no problems with remembering things ☐

I have slight problems with remembering things ☐

I have moderate problems with remembering things ☐

I have severe problems with remembering things ☐

I am unable to remember things ☐

COGNITION (option 2)

I have no problems in thinking clearly ☐

I have slight problems in thinking clearly ☐

I have moderate problems in thinking clearly ☐

I have severe problems in thinking clearly ☐

I am unable to think clearly ☐

COGNITION (option 3)

I have no problems working things out ☐

I have slight problems in working things out ☐

I have moderate problems in working things out ☐

I have severe problems in working things out ☐

I am unable to work things out ☐

ENERGY (option 1)

I have no problems with my energy levels ☐

I have slight problems with my energy levels ☐

I have moderate problems with my energy levels ☐

I have severe problems with my energy levels ☐

I have extreme problems with my energy levels ☐

ENERGY (option 2)

I am not tired ☐

I am slightly tired ☐

I am moderately tired ☐

I am severely tired ☐

I am extremely tired ☐

ENERGY (option 3)

I am not worn out ☐

I am slightly worn out ☐

I am moderately worn out ☐

I am severely worn out ☐

I am extremely worn out ☐

SLEEP

I have no problems sleeping ☐

I have slight problems sleeping ☐

I have moderate problems sleeping ☐

I have severe problems sleeping ☐

I have extreme problems sleeping ☐

Table 6.2 – Frequencies of participants' background characteristics for the two focus groups

	Frequencies focus group I	***Frequencies focus group II***
Age		
18-25	2	2
25-34	3	0
35-44	0	2
45-54	0	0
55+	0	2
Ethnicity		
White	5	5
Black, African or Caribbean	0	1
Gender		
Male	0	3
Female	5	3
Higher education achieved		
O-levels/ GCSE	0	1
A-levels	1	1
First degree	1	1
University higher degree	3	3
Professional activity		
Student	2	3
Employed or self employed	2	2
Student and employed or self employed	1	0
Retired	0	1
Marital status		
Married or living with a partner	3	4
Single or never married	2	2

Has the participant or someone close to her ever experienced a serious illness?		
Family or friends	2	0
No	3	0
Myself	0	6
Does the participant have children?		
Yes	0	0
No	5	6
Is the participant religious?		
Yes	1	4
No	4	2
Has the participant any previous experience with the EQ-5D?		
Yes	0	1
No	5	5

Table 6.3 – Additional bolt-on variants for the patients' focus group

VISION

I have no problems seeing while using any visual aid I need ☐

I have slight problems seeing while using any visual aid I need ☐

I have moderate problems seeing while using any visual aid I need ☐

I have severe problems seeing while using any visual aid I need ☐

I am unable to see ☐

RELATIONSHIPS

I have no problems with the relationships I care for ☐

I have slight problems with the relationships I care for ☐

I have moderate problems with the relationships I care for ☐

I have severe problems with the relationships I care for ☐

I am unable to have the relationships I care for ☐

LIFE SATISFACTION

I am extremely satisfied with my life ☐

I am very satisfied with my life ☐

I am satisfied with my life ☐

I am quite satisfied with my life ☐

I am not satisfied with my life ☐

Table 6.4 - Final descriptors and labels for the 8 bolt-on dimensions

HEARING	
I have no problems hearing	☐
I have slight problems hearing	☐
I have moderate problems hearing	☐
I have severe problems hearing	☐
I am unable to hear	☐

SLEEP	
I have no problems sleeping	☐
I have slight problems sleeping	☐
I have moderate problems sleeping	☐
I have severe problems sleeping	☐
I have extreme problems sleeping	☐

RELATIONSHIPS	
I have no problems with my social relationships	☐
I have slight problems with my social relationships	☐
I have moderate problems with my social relationships	☐
I have severe problems with my social relationships	☐
I am unable to have social relationships	☐

ENERGY

I have no problems with my energy levels ☐

I have slight problems with my energy levels ☐

I have moderate problems with my energy levels ☐

I have severe problems with my energy levels ☐

I have extreme problems with my energy levels ☐

COGNITION

I have no problems with remembering things ☐

I have slight problems with remembering things ☐

I have moderate problems with remembering things ☐

I have severe problems with remembering things ☐

I am unable to remember things ☐

LIFE SATISFACTION

I am extremely satisfied with my life ☐

I am very satisfied with my life ☐

I am moderately satisfied with my life ☐

I am slightly satisfied with my life ☐

I am not satisfied with my life ☐

SPEECH

I have no problems with my speech ☐

I have slight problems with my speech ☐

I have moderate problems with my speech ☐

I have severe problems with my speech ☐

I am unable to speak ☐

VISION (while using any visual aid you might need)

I have no problems seeing ☐

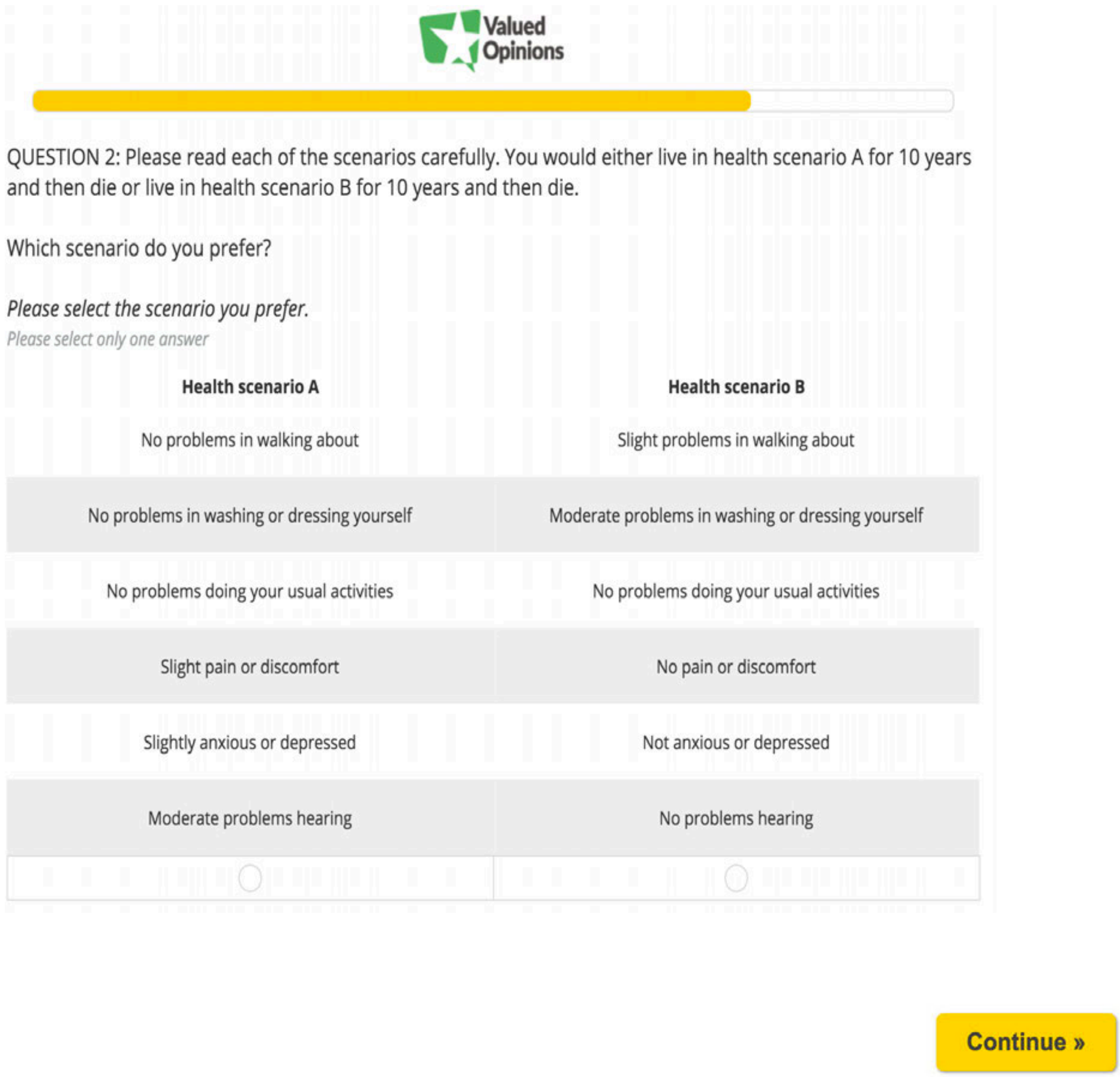
I have slight problems seeing ☐

I have moderate problems seeing ☐

I have severe problems seeing ☐

I am unable to see ☐

Figure 6.1 – Example of pairwise comparison as presented to responders


Valued Opinions

QUESTION 2: Please read each of the scenarios carefully. You would either live in health scenario A for 10 years and then die or live in health scenario B for 10 years and then die.

Which scenario do you prefer?

Please select the scenario you prefer.

Please select only one answer

Health scenario A	Health scenario B
No problems in walking about	Slight problems in walking about
No problems in washing or dressing yourself	Moderate problems in washing or dressing yourself
No problems doing your usual activities	No problems doing your usual activities
Slight pain or discomfort	No pain or discomfort
Slightly anxious or depressed	Not anxious or depressed
Moderate problems hearing	No problems hearing

Continue »

Table 6.5 – Background characteristics of the survey sample

		Frequencies (Percentages)
Gender	Female	520 (50%)
	Male	520 (50%)
Age	18-24	99 (10%)
	25-34	216 (21%)
	35-44	220 (21%)
	45-54	227 (22%)
	55-64	160 (15%)
	65 +	118 (11%)
Status	Single	374 (36%)
	Married	597 (57%)
	Separated	42 (4%)
	Widowed	20 (2%)
	Prefer not to say	7 (1%)
Education	O-level / GCSE	246 (24%)
	A-level	198 (19%)
	Diploma	105 (10%)
	First Degree	286 (28%)
	Postgraduate Degree	163 (16%)
	Other	42 (4%)
Employment	In employment	650 (63%)
	Retired	149 (14%)
	Homemaker	71 (7%)
	Student	71 (7%)
	Seeking work	17 (2%)
	Unemployed	34 (3%)
	Long term sick	42 (4%)
	Other	6 (1%)
Children	Yes	545 (52%)

	No	495 (48%)
Religion	Yes	300 (29%)
	No	705 (68%)
	Prefer not to say	35 (3%)

Table 6.6 – Self-reported health status of the survey sample

Self-reported health status dimension	*Level of dimensions*	*Frequencies (percentages)*
Mobility	Level 1	802 (77%)
	Level 2	159 (15%)
	Level 3	52 (5%)
	Level 4	26 (3%)
	Level 5	1 (0%)
Self-care	Level 1	940 (90%)
	Level 2	67 (6%)
	Level 3	29 (3%)
	Level 4	4 (0%)
	Level 5	0 (0%)
Usual activities	Level 1	795 (76%)
	Level 2	151 (15%)
	Level 3	65 (6%)
	Level 4	22 (2%)
	Level 5	7 (1%)
Pain / Discomfort	Level 1	533 (50%)
	Level 2	362 (35%)
	Level 3	111 (11%)
	Level 4	28 (3%)
	Level 5	6 (1%)
Anxiety/ Depression	Level 1	565 (56%)
	Level 2	266 (23%)
	Level 3	134 (13%)
	Level 4	54 (5%)
	Level 5	21 (2%)
Hearing	Level 1	850 (82%)
	Level 2	141 (14%)
	Level 3	41 (4%)
	Level 4	8 (1%)
	Level 5	0 (0%)

Sleep	Level 1	487 (47%)
	Level 2	338 (33%)
	Level 3	137 (13%)
	Level 4	65 (6%)
	Level 5	13 (1%)
Cognition	Level 1	658 (63%)
	Level 2	311 (30%)
	Level 3	63 (6%)
	Level 4	7 (1%)
	Level 5	1 (0%)
Energy	Level 1	447 (43%)
	Level 2	365 (35%)
	Level 3	163 (16%)
	Level 4	52 (5%)
	Level 5	13 (1%)
Relationships	Level 1	729 (70%)
	Level 2	204 (20%)
	Level 3	66 (6%)
	Level 4	30 (3%)
	Level 5	11 (1%)

Table 6.7 – Frequencies and percentages of choice options for pairwise choices with and without bolt-on

Bolt-on	*Bolt-on level*	*Frequencies (percentages)*					
		Health state pair 1		*Health state pair 2*		*Health state pair 3*	
		Choice A 11122	*Choice B* 23111	*Choice A* 52211	*Choice B* 11325	*Choice A* 33142	*Choice B* 34333
No bolt-on	/	106 (63.9%)	62 (36.1%)	90 (51.7%)	84 (48.3%)	118 (70.6%)	49 (29.4%)
Hearing	Level 1	94 (56.3%)	73 (46.7%)	78 (53.3%)	89 (46.7%)	125 (70.6%)	52 (29.4%)
	Level 3	68 (39.1%)	106 (60.9%)	84 (48.3%)	90 (51.7%)	79 (43.9%)	101 (56.1%)
	Level 5	18 (10.7%)	150 (89.3%)	37 (20.6%)	143 (79.4%)	35 (20.1%)	139 (79.9%)
Sleeping	Level 1	98 (56.3%)	76 (43.7%)	98 (58.3%)	76 (43.7%)	107 (54.1%)	60 (35.9%)
	Level 3	70 (38.9%)	110 (61.1%)	70 (41.7%)	98 (59.3%)	88 (50.6%)	86 (49.4%)
	Level 5	58 (32.8%)	119 (67.2%)	53 (31.7%)	114 (68.3%)	80 (45.2%)	97 (54.8%)
Cognition	Level 1	96 (54.2%)	81 (45.8%)	79 (43.9%)	101 (56.1%)	107 (64.1%)	60 (35.9%)
	Level 3	60 (34.5%)	114 (65.5%)	71 (40.8%)	103 (59.2%)	61 (36.3%)	107 (63.7%)
	Level 5	31 (17.8%)	143 (82.2%)	61 (34.5%)	113 (65.5%)	44 (25.3%)	130 (74.7%)
Energy	Level 1	116 (65.4%)	64 (35.6%)	92 (52.0%)	85 (48.0%)	113 (63.8%)	64 (36.2%)
	Level 3	81 (48.5%)	86 (51.5%)	93 (53.5%)	81 (46.5%)	89 (51.1%)	85 (48.9%)
	Level 5	35 (20.8%)	133 (79.2%)	59 (32.8%)	121 (67.2%)	57 (33.9%)	111 (66.1%)
Relationships	Level 1	106 (58.9%)	74 (41.1%)	81 (45.8%)	96 (54.2%)	89 (53.0%)	79 (47.0%)
	Level 3	78 (46.7%)	89 (53.3%)	65 (38.7%)	103 (61.3%)	94 (52.2%)	86 (47.8%)
	Level 5	42 (24.1%)	132 (75.9%)	51 (29.3%)	123 (70.7%)	48 (27.6%)	126 (72.4%)

Note: Bolt-ons at a moderate (level 3) and severe (level 5) level were added to choice option A. Choice option B always presented bolt-ons at level 1

231

Table 6.8 – Responders' preference for option A for bolt-ons at level 3 ranked by size (from largest to smallest impact)

Health state pair 1	
Bolt-on	*Percentages*
Cognition	60 (34.5%)
Sleeping	70 (38.9%)
Hearing	68 (39.1%)
Relationships	78 (46.7%)
Energy	81 (48.5%)

Health state pair 2	
Bolt-on	*Percentages*
Relationships	65 (38.7%)
Cognition	71 (40.8%)
Sleeping	70 (41.7%)
Hearing	84 (48.3%)
Energy	93 (53.5%)

Health state pair 3	
Bolt-on	*Percentages*
Cognition	61 (36.3%)
Hearing	79 (43.9%)
Sleeping	88 (50.6%)
Energy	89 (51.1%)
Relationships	94 (52.2%)

Note: Comparisons are based on the ranking of the bolt-ons and not the difference in percentage between them.

Table 6.9 – Responders' preference for option A for bolt-ons at level 5 ranked by size (from largest to smallest impact)

Health state pair 1	
Bolt-on	*Percentages*
Hearing	18 (10.7%)
Cognition	31 (17.8%)
Energy	35 (20.8%)
Relationships	42 (24.1%)
Sleeping	58 (32.8%)

Health state pair 2	
Bolt-on	*Percentages*
Hearing	37 (20.6%)
Relationships	51 (29.3%)
Sleeping	53 (31.7%)
Energy	59 (32.8%)
Cognition	61 (34.5%)

Health state pair 3	
Bolt-on	*Percentages*
Hearing	35 (20.1%)
Cognition	44 (25.3%)
Relationships	48 (27.6%)
Energy	57 (33.9%)
Sleeping	80 (45.2%)

Note: Comparisons are based on the ranking of the bolt-ons and not the difference in percentage between them.

Chapter 7
Discussion and Conclusions

7.1 Introduction

The aim of this thesis was to explore methods to identify and select bolt-on dimensions for GPBMs. Once bolt-ons have been identified and selected, health state utility values for the GPBM together with one or more bolt-ons can be obtained using preference elicitation techniques. Health state utility values can be subsequently used in economic evaluations of healthcare interventions. Given the limited resources available, it would have been unfeasible to conduct this research on all GPBMs. For this reason, the EQ-5D was chosen as a case study.

A number of objectives needed to be achieved to address this thesis aim. First, it was necessary to review the literature on the psychometric evidence of GPBMs to justify why bolt-on research might be needed. Second, it was important to explore methods for identifying bolt-ons. Third, it was necessary to explore methods for selecting bolt-ons.

Exploring methods implied assessing their feasibility for identifying and selecting bolt-ons, comparing different techniques within the methods used (if relevant), understanding whether the methods were useful to inform decisions, considering the difficulties in applying them and recommending ways to improve them if possible. These objectives were met through a number of studies that have been reported in this thesis.

This final chapter presents the key findings of the thesis, its contributions to the existing knowledge, its limitations and provides a list of recommendations for future research.

7.2 Key findings

This thesis and the studies it includes offer a number of interesting findings. These are presented in the next sections.

7.2.1 Overview of systematic reviews

The study included in Chapter 3 reviewed evidence on the validity and responsiveness of the 5 most used GPBMs across disease areas and conditions using an overview of systematic reviews. This study was motivated by the fact that the chosen GPBM should be appropriate for the group of patients being examined in the evaluation and should be able to detect meaningful change. If the investigated GPBM reports problems of validity and responsiveness, this justifies further investigation into the possibility of adding bolt-ons. The study generated 4 important sets of findings.

One set of findings is directly linked with the objective of justifying the need for bolt-on research by assessing the performance of the 5 most used GPBMs across disease areas and conditions. The study showed that when evidence is available, it often supports the validity and responsiveness of the investigated measures. For example, the 3L version of the EQ-5D seemed valid and responsive in many conditions including respiratory, genitourinary, endocrine, nutritional and metabolic diseases. Despite this, the overview also found that all GPBMs had some problems of validity and responsiveness in at least one condition or disease area. For instance, the 3L version of the EQ-5D performed poorly in hearing impairments, multiple sclerosis and a number of mental health conditions. These problems limit the appropriateness of using a single measure in all assessments and highlight the relevance of bolt-on research.

Another set of findings regards the size and coverage of the available evidence included in the literature reviews. The overview found that the vast majority of studies investigated the EQ-5D, with significantly fewer studies being conducted for the remaining GPBMs. Even for the EQ-5D, many conditions were not covered, and often only one indicator of validity and responsiveness was used e.g. known group method. These findings highlight that the evidence produced to date is not comprehensive, and this pose doubts on its usefulness to compare GPBMs.

A third set of findings regards the quality, nature and reporting of evidence in literature reviews of psychometric evidence. The review found that most of the literature reviews were poor in quality, reported studies that often used indicators of differences or change that were inappropriate to assess GPBMs e.g. clinical indicators rather than patient reported outcomes and lacked consistency in presenting the thresholds used to assess the findings, the statistical significance of these findings and confidence intervals for the estimates. This further limits the usefulness of this type of evidence to assess GPBMs.

Finally, some conclusions can be drawn in regard to whether reviews can be used as a method to identify bolt-ons, which was the case in some of the bolt-on studies identified in the literature (e.g. Yang et al, 2015; Hogendoorn et al, 2016). By comparing different reviews it emerged that the evidence they include does not enable the identification of specific dimensions missing from the investigated GPBM descriptive system, or dimensions that might be relevant across multiple conditions. This is due to two main reasons. First, literature reviews generally focus on utility scores rather than unscored dimensions, and this does not allow the assessment of the comparative performance of the individual dimensions in different populations. Second, issues in the quality, conduct and reporting of literature reviews discussed earlier, such as comparisons of GPBMs with clinical indicators rather than patient reported outcomes and lack of clarity in the thresholds employed, undermine the possibility of informing decisions for identifying bolt-ons based on this evidence.

Although, as currently used, literature reviews do not appear a feasible method to identify bolt-ons, they still have some value in bolt-on research as they allow the identification of conditions or disease areas where the investigated GPBM reports poor validity and responsiveness. They can therefore be used to signal the potential need for conducting bolt-on research in that area. If literature reviews are used for this latter purpose, these have to adhere to high standard of methodological quality, and use tests that are appropriate for GPBMs assessment. However, even in those cases their usefulness might be limited by the scarcity of psychometric evidence for

GPBMs other than EQ-5D, and the scarcity of evidence for the EQ-5D in many conditions.

7.2.2 Factor analysis techniques (EFA, PCA, CFA and SEM)

Chapter 4 examined the use of factor analysis for identifying bolt-ons. This study was motivated by the objective of exploring methods to identify bolt-on dimensions. It investigated the use of two exploratory techniques, PCA and EFA, comparing different methods of extraction, selection and rotation of components and factors. It extended the most interpretable exploratory model i.e. categorical PCA into a confirmatory model i.e. CFA. It then assessed differences in terms of bolt-ons identified between PCA, EFA and CFA. The chapter also examined the use of SEM to better understand the relationship between factors and to avoid double counting of bolt-ons. The factor analysis study generated a number of important findings that address the aim and objectives of this thesis.

The study showed that factor analysis is a feasible method to identify bolt-on dimensions, whether these are specific bolt-ons i.e. bolt-ons that can improve the measure in a specific condition or generic bolt-ons i.e. bolt-ons that can improve the measure across multiple diseases. As factor analysis relies on the identification of the latent structure to which a set of items taken from multiple measures relate, the technique can also be used to identify bolt-ons that measure aspects of health already covered by one or more dimensions included in the parent measure e.g. if the objective is improving the performance of the EQ-5D in mental health, the technique could be used to expand its descriptive system with bolt-ons measuring other aspects of psychological symptoms such as calmness and agitation, or self-harm. It should be noticed that the feasibility of this method largely depends on the availability of a dataset that comprehensively covers the domains space of interest. This latter point should always be ensured if factor analysis is used for the purpose of bolt-on identification.

In comparing different techniques, the study found that PCA and EFA are equivalent exploratory methods to investigate the latent structure to which the selected measures relate. More importantly, little differences exist in terms of bolt-ons identified using PCA, EFA and CFA. Notwithstanding these similarities, both PCA and EFA identified 8 items as bolt-ons that were not confirmed using CFA. In some of these cases, differences in the bolt-ons identified originated from items loading on different factors in CFA compared to their respective components or factors in PCA and EFA. These findings are useful in informing the identification of bolt-ons. They show that if the interest is in identifying factors as bolt-on dimensions, these three techniques appear equivalent and interchangeable. However, if the interest is in adapting items into bolt-on dimensions, CFA should always be used in conjunction with the other techniques as differences exist in the items identified as bolt-ons. The study also found that SEM might be, from a strictly empirical point of view, a feasible method to understand the relationship between factors and reduce the risk of double counting of bolt-ons. However, it also concluded that this technique reports similar goodness of fit indexes for different models and for this reason should only be used if agreement exists on a conceptual model of health and health measures. The issue of whether agreement can be reached on a conceptual model of health and health measures is beyond the scope of this thesis, but represents a major concern before employing SEM.

The findings of the factor analysis study are also useful in informing a discussion on how these techniques should be used. Differences were noted in the items identified as bolt-ons when using different thresholds. For example, in PCA the use of a threshold of 0.45 for item loadings led to the identification of the item SF-6D role as a candidate bolt-on. Using another commonly employed threshold i.e. 0.3 would have resulted in the same item being considered as mainly loading on physical functioning and therefore being excluded from the list of bolt-ons. As the choice of which threshold should be used for considering item loadings relevant has a potentially large impact in terms of items identified as bolt-on dimensions, as well as on the interpretation of the factor to which items relate, this aspect should be given considerable attention when using factor analysis methods for identifying bolt-ons. Some elements might aid this choice, among which the interpretability of factors based on different thresholds and the interpretability and consistency of items loading on factors.

Another important finding of this study is the fact that little difference existed in terms of results obtained when comparing methods specifically geared for categorical data and methods treating ordinal categorical variable as continuous. This suggests that although using techniques specifically geared to the data used would be preferable, use of the standard methods does not substantially alter conclusions.

In addition to the findings directly linked to the objective of this thesis, the factor analysis study generated two additional relevant results. First of all, it identified 6 factors that are not covered by the 5 dimensions of the 5L version of the EQ-5D. These are relationships, speech/cognition, energy/sleep, hearing, satisfaction and vision. These factors can be developed into bolt-ons for use in the EQ-5D-5L, though there needs to be consideration of other issues such as how they fit the conceptual framework and what should be the scope of the measure. Second, it also identified a list of 37 items related to the above-mentioned factors that could be adapted into bolt-on dimensions for the EQ-5D, and 9 additional items not related to any of the factor identified that might be adapted into bolt-ons.

7.2.3 Selecting bolt-ons using their ability to predict HRQoL

Having found a useful method for identifying bolt-ons, Chapter 5 introduced the next stage of how to select the best dimensions i.e. factors and items identified from the factors analysis. This study was motivated by the fact that for each bolt-on option the new descriptive system needs to undergo a complete re-evaluation, a process that can be costly and complex. Therefore, selection is important to ensure that bolt-ons are relevant. For this purpose, the chapter examined the possibility of selecting bolt-ons based on their ability to detect differences in a proxy of HRQoL measured using the Health VAS. The chapter employed two tests based on linear regressions. In the first test, factors and items were compared in terms of their ability to explain HRQoL over and above the EQ-5D. In the second test, factors and items were compared in terms of their ability to explain HRQoL over and above the EQ-5D for the coefficients of 9 chronic conditions dummies. The study generated a number of important findings related to the aims and objectives of this thesis.

The first test appeared a feasible technique to discriminate between bolt-ons, as it was able to detected systematic differences in factors' and items' ability to predict variations in HRQoL. Results between regressions based upon factors and items were generally consistent, as items loading on factors registering large coefficients generally reported large coefficients, and items loading on factors reporting small coefficients generally reported small coefficients. This evidence can be used to inform the selection of bolt-ons, as a large coefficient for a factor or item signals that that factor or item might improve the construct validity of the investigated GPBM. Differences were also noted between items loading on the same factor, and these appeared related to the aspects of health measured by them. Additionally, this information can be used not only to select the most relevant items, but also to inform choices on how to develop bolt-ons from factors. For example, if systematic differences in terms of coefficients are found between items measuring similar aspects of health but loading on the same factor as found for speech, cognition, energy and sleep, this information can be used to decide whether to develop a single bolt-on from a factor, or multiple bolt-on options. The first test also registered reductions in the coefficients of the core dimensions of the investigated GPBM when some items were added to the model. These reductions were easily interpretable and appeared to signal an interaction between the items and the core dimensions of the reference GPBM. This is relevant for selecting bolt-ons, as the choice of which should be developed or adapted should aim to minimise the overlap between dimensions i.e. minimising the impact of factors and items on the coefficients of the core dimensions. It is worth mentioning that despite being informative, results from regressions should not be considered a definitive test. Other information is also important for selecting bolt-ons such as the scope for adding a bolt-on dimension or the strength of preferences for that bolt-on.

Results from the second test were sometimes difficult to interpret, as discrepancies were noticed between items and factors regressions. More specifically, while reductions in beta coefficients generally occurred for chronic conditions theoretically related to the aspects of health measured by the items, this was not always the case for factors. The presence of discrepancies in the results of factors and items regressions using the second test poses some doubts on the feasibility of this method for bolt-ons selection. For this reason, further investigation is needed before using this technique.

In addition to the findings directly linked to the aims and objectives of this thesis, this study provides two additional sets of findings. A first set of findings regards the EQ-5D and its items' ability to predict HRQoL. The study showed that the EQ-5D-5L is generally able to explain variations in HRQoL as measured by the Health VAS. However, it also found an interaction between self care and usual activities, which may explain the lack of ability of self care to predict HRQoL in this study.

A second set of findings regards the behaviour of the factors and items tested and their selection as bolt-on dimensions. The study found that all the factors and most of the items tested were able to explain differences over and above the 5L version of the EQ-5D. Those factors and items represent important potential bolt-on options. Energy/sleep, relationships and satisfaction, and the items related to these factors, appeared better than speech/cognition, vision and hearing, and the items related to these factors, at predicting HRQoL. Moreover, coefficients for the items related to energy and cognition were consistently higher than coefficients for the items related to sleep and speech. This suggests that energy/sleep, relationships and satisfaction should be preferred to the remaining factors for bolt-on development, and energy to sleep and cognition to speech if items were to be adapted into bolt-ons.

7.2.4 Selecting bolt-ons using their impact on preferences for health states

Chapter 6 reported a study that tested an alternative approach for selecting bolt-ons, namely using preferences elicited from pairwise choices. The study was structured into two phases. Firstly, bolt-on dimensions were developed from the identified factors and tested in terms of their relevance and face validity using two focus groups. Subsequently, a pairwise choice survey was conducted. The survey used a subset of the developed bolt-ons (5 bolt-ons chosen based on the results of Chapter 5 and evidence from previously published literature) and examined whether their addition at level 1, 3 and 5 to 3 selected health state pairs caused a switch in preferences between the states. It compared the magnitude of this switch between bolt-ons and within bolt-ons i.e. level 1 versus 3 and level 1 versus 5, assuming that the larger was the switch, the greater would be the impact of the bolt-on in a valuation study.

The pairwise choices study generated important findings that address the aims and objectives of this thesis. First, it showed that using preferences for pairwise choices to discriminate between bolt-ons is a feasible method. Each of the 5 bolt-ons tested reduced the number of responders choosing the health state associated to the bolt-ons at level 3 and 5, while as expected level 1 bolt-ons did not affect preferences over pairwise choices in a statistically significant way. This evidence can be used to inform on the relevance of the bolt-on dimensions, as large switches in preferences for a bolt-on signal that this would likely have a large impact in a valuation study, and small switches a small impact. Differences were found between bolt-ons in terms of their impact on preferences for the same pairs of health states when these were added at level 3, and at level 5. This information can be used to discriminate between bolt-ons, as if when comparing two bolt-ons at the same level one consistently registers larger switches in preferences than the other, this suggests that the aspect of health convered by the first bolt-on is considered more important by responders.

This study found differences in the switches in preferences caused by the same bolt-on at the same level if added to different pairs of health states. This is an important finding that requires careful consideration when using this method to select between bolt-ons. As the impact of a bolt-on on preferences depends on the health states to which it is added, it is advisable to use a large number of health states to inform on the best bolt-on option, as this gives a full assessment of the overall impact that a bolt-on would have in a valuation study.

Moreover, this study found that the comparative switch in preference between bolt-ons differed depending on whether level 3 or level 5 were compared e.g. at level 3 cognition generally produced the larger switch in preferences, while at level 5 hearing. An implication of this finding for the purpose of selecting bolt-ons is that a decision rule is required in terms of what is considered more important. For example, the interest could be in appending the bolt-on with the largest impact on preferences for health states, in which case only bolt-ons at level 5 would be used for comparisons. Alternatively, the bolt-on having on average the largest impact i.e. the bolt-on reporting the largest switches for each of its levels could be selected. This decision is beyond the scope of this thesis but is an important aspect requiring consideration for an appropriate use of this method.

In addition to the findings specifically related to the aims and objectives of this thesis, the pairwise comparison study provided relevant information for the specific bolt-ons tested and the EQ-5D. It showed that cognition, sleep, hearing, energy and relationships have an impact on preferences for the 5L version of the EQ-5D. When a level 3 was added to existing pairs of health states, cognition produced the largest switch in preferences while energy had the smallest impact. Sleeping and hearing reported similar impacts while relationships showed an inconsistent pattern depending on the health state pairs. By contrast, when a level 5 was added, hearing registered the largest impact on preferences, cognition, energy and sleep registered a similar impact and sleep the smallest impact.

7.3 Contribution to the existing knowledge

The findings of this thesis and the individual studies it includes offer important contributions to the existing knowledge and literature of the bolt-on area of research. Specific contributions of the individual studies are listed below, followed by a discussion on the broader contribution of this research program.

7.3.1 Overview of systematic reviews

The overview of systematic reviews reported in Chapter 3 is the first study that summarizes evidence on the validity and responsiveness of the five most used GPBMs across all disease areas and conditions. This contributes to the knowledge by providing a useful overall assessment of the state of the art in terms of GPBMs performance and the overall quality and quantity of this evidence. While some conclusions can be drawn on the actual performance of GPBMs, the overview shows how incomplete and poor much of the evidence is. The overview also contributes to the current knowledge by showing that the way in which evidence is currently summarized in literature reviews of psychometric evidence is of limited usefulness for identifying bolt-ons.

7.3.2 Factor analysis techniques (EFA, PCA, CFA and SEM)

The factor analysis study presented in chapter 4 compared 4 techniques to identify bolt-ons. None of these techniques had been tested for the purpose of identifying bolt-ons prior to the beginning of this thesis, and only one of them, PCA, was employed for selecting between alternative bolt-on items while this research program was being implemented (Hogendoorn et al, 2016). The study demonstrates that factor analysis is a feasible technique for identifying bolt-ons. It allowed the identification of a set of factors, and items related to them, that are reasonable potential additions to the EQ-5D. The results of this study were generally consistent with exploratory work conducted in the field, confirming in some cases the relevance of some bolt-ons such as energy and sleep (Gudex et al, 1991; Yang et al, 2015), and in others identifying new bolt-on options e.g. relationships. This represents an important advancement in the field, as it provides a tool to help identify bolt-ons in a structured way using quantitative data. By comparing the latent constructs to which the most commonly used GPBMs relate in a comprehensive and systematic way, this method avoids the arbitrariness of other techniques e.g. expert opinion, and simultaneously ensures that all aspects of health that might be relevant are considered. The factor analysis study also shows that despite the fact that most of the items identified as candidate bolt-ons using exploratory techniques are confirmed using confirmatory methods, the practise of using only exploratory factor analysis as done in Hogendoorn et al (2016) may lead to an incorrect identification of items as relevant. This is because the component on which some of the items load using PCA might differ from the factor on which the same items load using CFA. Finally, the study shows that SEM can be used to confirm how factors are related in the presence of an agreed conceptual model of health. However, the issue of whether it is actually possible to reach agreement on a conceptual model of health and health measures is of crucial importance for the correct use of this technique. In the current study, agreement of the research group has been reached on one model only for the purpose of assessing, methodologically, the possibility of using this technique. The issue of whether a broader agreement can be found is a major concern before this technique can successfully be used.

7.3.3 Selecting bolt-ons using their ability to predict HRQoL

The study presented in chapter 5 represents the most comprehensive effort ever conducted in comparing the ability of different bolt-ons to add to the explanatory power of the EQ-5D. Its findings add to those of previously conducted studies (e.g. Jelsma and Maart, 2015; Perneger and Curvoisier, 2011) in showing that a number of additional dimensions may help improving the explanatory power of the EQ-5D. Differently from these studies, a large number of bolt-on dimensions and items were tested (each of the two studies investigate the addition of only 5 items), and these were identified systematically through the use of quantitative techniques. In doing this, the study also provides new evidence on the comparative contribution of different items and factors in predicting differences in HRQoL as measured by the Health VAS. It showed that numerous items and factors might improve the construct validity of the EQ-5D, with some of them having a substantially larger impact than others e.g. relationships compared to hearing. It also highlights that some items might interact with the dimensions included in the core descriptive system of the EQ-5D, signalling that items with the least interactions should be the preferred options if adapted into bolt-ons.

7.3.4 Selecting bolt-ons using their impact on preferences for health states

The study presented in chapter 6 is the first methodological study that uses pairwise choices to assess the impact of bolt-on dimensions in terms of switches in preferences between health states. This technique has never been used for this purpose before, and in that it contributes to the body of knowledge providing a useful method for discriminating between bolt-ons. The study also informs on the fact that the switch in preferences for different health states is not constant across bolt-ons at different levels of the dimensions. This implies that the choice between bolt-ons might require a decision rule on whether it is considered more important for a bolt-on to generate the largest impact on preferences in absolute terms i.e. the greatest switch in preferences at level 5, or for a bolt-on to generate, on average, the largest impact i.e. a bolt-on that causes on average the largest switch.

7.3.5 General contributions to knowledge

This research project represents the first attempt to systematically explore methods for identifying and selecting bolt-on dimensions. This is an extremely relevant contribution considering that the bolt-on studies conducted untill now tended to identify and select bolt-ons using methods in an unstructured way (e.g. Yang et al, 2014; Gudex, 1991). In exploring these methods, some of which used by the above mentioned studies e.g. literature reviews of psychometric evidence (e.g. Yang et al, 2015; Swinburn et al, 2013), this research program generated an important evidence base that helps in understanding which methods can be used for identifying and selecting bolt-ons and what information they provide.

More specifically, despite the fact that the overview of systematic reviews and the factor analysis study generated different types of evidence, they often pointed in a similar direction in terms of identification of potential areas for bolt-ons. For example, the overview of reviews identified problems for the EQ-5D in hearing impairments and visual disorders, and the factor analysis found hearing and vision as bolt-ons missing from the EQ-5D descriptive system. Similarly, the overview of systematic reviews found problems in the EQ-5D psychometric characteristics in cardiovascular diseases, which are often associated with tiredness, fatigue and impaired thinking (American Hearth Association, 2017). These aspects might be captured by bolt-ons identified in the factor analysis such as cognition, energy and sleep. This suggests that there might be a benefit in using these two types of evidence complementary, triangulating their results, for the purpose of identifying bolt-ons. This helps to inform current practice and sets a benchmark against which to compare the quality of the future research in the field.

By contrast, the study examining the selection of bolt-ons based on their ability to predict differences in HRQoL and the study examining the selection of bolt-ons based on their impact on preferences for pairwise choices generated different results. For example, while energy/sleep appeared to be the most important factor in the regression study, energy ranked 5th when tested for level 3 and 4th when tested for level 5 in the pairwise choice study. Similarly, while the factor speech/cognition ranked 4th in the regression study, cognition ranked 1st when tested at level 3 and 2nd when tested at level 5 using pairwise choices. There are multiple explanations for these differences, among which limitation of the Health VAS (see next section), and the fact that the Health VAS reflects patients views while preferences for pairwise choices were obtained from a general population sample in which preferences may not be as well informed i.e. asymmetry of information between patients completing a VAS and general public responders who are asked to choose between health states (e.g. Karimi et al, 2017). Due to these differences, the choice of which of these techniques should be used is not obvious and normative judgemnets may be required.

While not the primary aim of this thesis, the findings included here provide important contributions to the knowledge and research for the EQ-5D. Consistently with previous research (e.g. Brazier et al, 2014; Longworth et al, 2014) this research program shows that the EQ-5D is a generally valid and responsive measure in numerous conditions. However, similarly to other studies (e.g. Longworth et al, 2014), it also shows that the measure might report problems of validity and responsiveness in some disease areas. A possible explanation for this lack of validity and responsiveness might be the fact that the measure covers only some of the important HRQoL constructs. More specifically, this thesis shows that dimensions like vision, hearing, speech, relationships, life satisfaction, cognition, energy and sleep might be important to describe HRQoL and are not covered by the EQ-5D. These findings extend on the results of previous qualitative research on the aspects of health missing from the EQ-5D (e.g. Saha et al, 2016; Devlin et al, 2004) providing the most exhaustive (although not complete) list of candidate bolt-ons for this GPBM. On these grounds, these findings can be used by the EuroQol group to set the agenda for future developments of the measure.

Finally, this thesis generated a number of valuable scientific outcomes. Two of the studies included in this thesis i.e. the overview of reviews of chapter 3 and the factor analysis study of chapter 4 have been published in the European Journal of Health Economics and in Value in Health (Finch et al, 2017; Finch et al, 2017). The research reported in this thesis has been disseminated in numerous international conferences, among which the 2015 International Society for Pharmacoeconomics and Outcome Research (ISPOR) 18th Annual European Congress, the 2016 Health Economics Study Group (HESG) Summer Conference, the 2016 33rd EuroQol Plenary meeting, the 2017 2nd EuroQol Academy Meeting, the 2017 2nd Advances in Patient Reported Outcomes Conference and the 2017 34th EuroQol plenary meeting. The study presented in Chapter 6 has been funded through an award obtained from the EuroQol group in date 31st of March 2017 following peer review assessment. Two papers based on the studies of chapter 5 and chapter 6 will be submitted soon.

7.4 Limitations

This research program and the studies it includes have a number of limitations that need to be mentioned.

This thesis represents the broadest effort ever conducted to examine methods for bolt-ons identification and selection. Nevertheless, the list of methods examined here is not exhaustive and other techniques might be useful complements to the ones investigated. For example, interviews and focus groups with members of the general public, patients and clinicians might be used to identify areas of health relevant to describe HRQoL, or specific dimensions missing from a selected GPBM. Similarly, selection between items related to the same factor could have been examined using item response theory, or standard psychometric analysis of items performance. Instead of the regression analysis and the pairwise choice experiment presented in Chapter 5 and 6, other methods could have informed the selection of bolt-ons. Among them, TTO or SG could have been used to assess the impact of bolt-ons on the existing health states of the EQ-5D. Alternatively, ranking and best worst scaling techniques (Brazier et al, 2009) could be used to inform on the relative importance of bolt-ons by means of directly comparing them with the 5 dimensions of the EQ-5D. Finally, also interviews on the ability of bolt-ons to detect important aspects of patients' lives can aid the selection of bolt-on dimensions.

This thesis investigated bolt-ons for the EQ-5D. The 5L version of the measure was employed in all the studies presented here except for the overview of reviews, where the 3L version was used. This choice was motivated by the fact that no systematic review studies had been conducted on the psychometric performance of the 5L version at the beginning of this research program (the valuation set for the 5L version was released only in 2016). However, as one possible reason for the lack of validity and responsiveness of a GPBM is its lack of a sufficient number of levels to detect changes in health, it cannot be assumed that to a poor performance of the EQ-5D-3L necessarily corresponds a poor performance of the EQ-5D-5L. Hence, inconsistencies in the comparison of results between the psychometric evidence and the factor analysis might exist.

A set of limitations of this thesis is related to the datasets used for the analysis of Chapter 4, 5 and 6. First of all, both datasets were based on online recruiting of individuals that have previously registered to a panel. This type of recruiting method might suffer from substantial self-selection bias. For example, severely ill patients might be excluded due to the impossibility of enrolling to the panel, or participants might be excluded as they do not have access to the internet. Moreover, this type of recruitment method is prone to poor quality of data, as it is not possible to verify that participants correctly understood the tasks presented to them. In order to account for this latter issue, quality checks were used in both datasets e.g. time taken in completing the questionnaires. Nevertheless, there is still a possibility that some poor quality data was used. Additionally, both datasets were cross-sectional, and this limits the possibility of assessing causality. This is particularly relevant for the study reported in Chapter 5, where bolt-ons were compared in terms of their ability to predict the Health VAS. A longitudinal dataset would have allowed endogeneity and time invariant variables to be controlled for, and this would have increased confidence in the findings being causal and not simply correlational (Wooldrigde, 2009). Finally, the Multi Instrument Comparison dataset used for the analysis of Chapter 4 and Chapter 5 is multinational. This raises the issue of the validity of instrument translations. Moreover, due to the large number of measures administered in the dataset, survey biases might have arose from the repetition of multiple items.

A final set of limitations of this thesis is related to the methods used and the variables included in the individual studies reported. The overview of reviews of Chapter 3 collected evidence from published and unpublished reviews of the literature. This implies that relevant evidence from studies not included in these reviews was not covered in the synthesis provided.

The study presented in Chapter 5 used linear regressions to predict the ability of bolt-ons to detect differences in HRQoL. Despite the fact that this choice was consistent with previously published literature (e.g. Jeelsma and Maart, 2015;

Perneger and Courvoisier, 2011), other models could have been investigated for the same purpose e.g. two limit tobit model. Moreover, this study used the Health VAS as a proxy of HRQoL. The EQ VAS has been extensively criticized as a measure of HRQoL e.g. non-health contextual factors influence the scores obtained, and despite the fact that the Health VAS differs from the EQ VAS, some of the limitations noted for the EQ VAS might aslo apply to this measure. In addition, as the Health VAS used in this study defines health in a very comprehensive way, it might be that some of the results obtained for factors or items would have been different using a VAS with a narrower definition of health. Finally, this study did not estimate interactions between factors and items and the chronic conditions despite it being theoretically plausible that in presence of a chronic health condition the amount of variance explained by a factor or item is larger. Interactions were initially calculated but were not reported as the large number of coefficients generated was of difficult to interpret. This choice might have reduced the validity of the second test performed.

Lastly, the study reported in Chapter 6 used pairwise choices to elicit preferences. This choice was based on the assumption that a large switch in preferences in the pairwise choice tasks would result in a large decrement in the coefficient associated to that level of the bolt-on in a valuation study. However, pairwise choices are recognized to generate utility values that differ from the TTO (e.g. Brazier et al, 2012; Stolk et al, 2010). For this reason, a valuation study using the TTO might generate health state values with decrements that are not consistent with the findings of this study.

This research program has explored methods to identify and select bolt-ons. However, it has not clarified how different methods should be employed according to the different objectives that might motivate the need for adding a bolt-on. For example, identification and selection of bolt-ons might be driven by the willingness of extending a GPBM into a broader measure of health, or it might be driven by the willingness of extending a GPBM into a broader measure that includes areas beyond health. These differences affect the criteria and the processes for identifying and selecting bolt-ons. Pursuing the first research goal would require the use of factor analysis to identify only factors related to health, while pursuing the second research goal would require factor analysis to be used to identify also factors related to well-being. This issue was not examined in this thesis. This highlights that the ultimate choice on bolt-ons is not just empirically driven, but also depends on the conceptual framework employed and the scope in terms of breath of measurement.

7.5 Future research

The findings and limitations of this thesis offer multiple suggestions for future research. These are listed below.

1. As studies included in this thesis were exploratory in nature, one set of recommendations for future research regards the possibility of further testing. In the factor analysis study, multiple rotation and extraction methods were compared with the objective of identifying the model of best fit. However, an in depth examination of the impact of these different choices on the identification of the latent structure and on the relationship between items and factors was not performed. This would be of interest, as it could help in clarifying whether the choices made should be recommended for all bolt-ons studies. In the pairwise choice study, only 5 of the 8 bolt-ons developed were used, and their impact was tested on only 3 of their 5 levels on a limited number of health states. This leaves space for future investigation of the impact of the remaining 3 bolt-ons on preferences for health states, and the impact of the other 2 levels of the 5 bolt-ons tested.

2. The studies included in this thesis investigated methods for identifying and selecting bolt-ons using the EQ-5D as a case study. However, these methods may also be useful for identifying bolt-ons for other outcome measures. Other GPBMs are commonly used in economic evaluations, and the possibility of identifying bolt-ons for the SF-6D and HUI3 is a valuable research area. Moreover, two of the methods used in this thesis, the factor analysis and the regression analysis, might also be relevant to identify and select bolt-ons for non-preference based measures.

3. Methods presented in this thesis could also be used to bolt-off dimensions. Dimensions to bolt-off could be identified as those loading on the same factor using factor analysis and selected using regression analysis or pairwise choices surveys. In the first case, the dimension with the lowest coefficient in terms of their ability to detect changes in HRQoL could be removed, while in the second case the dimension causing the smallest switch in preferences for alternative health states would be taken off.

4. This thesis noted differences in the results of the study examining the ability of bolt-ons to predict differences in HRQoL and the study examining the ability of bolt-ons to impact preferences for pairwise choices. One possible explanation for these differences may be driven by differences in experiences in that the Health VAS reflects patients experiences while pairwise choices are made by the general public, who may or may not have experiences of the problems described. For this reason, an interesting avenue for further research is represented by the investigation of whether differences in the results obtained originate from ill-informed preferences of general public responders. In order to assess this, an option would be running two pairwise choice experiments with members of the general public, providing in one of them more information on the experience of living in different health conditions. Differences in response choices for the two pairwise experiments might signal the presence of ill-informed preferences, or simply a poor measure of HRQoL.

5. Moreover, an interesting area for further research is represented by the investigation of how the techniques presented in this thesis should be combined together and with qualitative methods investigated in other studies (e.g. Saha et al, 2016) to inform the final selection of bolt-on dimensions. Different approaches could be used for this purpose, among which conducting expert groups or interviews with experts.

6. Furthermore, this thesis covered important aspects of the identification and selection of bolt-on dimensions, but did not investigate another area of research that is crucially important for a successful use of bolt-ons in economic evaluations, namely how to accommodate them in the value function of the GPBM. An initial attempt to estimate the impact of adding two bolt-ons simultaneously has been conducted by Swiburn and collegaues (2013) where interactions between dimensions in multi-linear models were estimated using linear least square and random effect models. Further research is recommended to clarify whether their findings are generalizable, and whether these models should be recommended for all valuations of GPBMs + bolt-ons.

7. Finally, for bolt-ons to be useful, these should be employed in economic evaluations. No study has been performed to date comparing the outcome of using the reference GPBM or the same GPBM + bolt-ons in the assessment of alternative treatments or interventions. This is an extremely interesting avenue for future research that would inform the likely impact of introducing bolt-ons in current practise.

7.6 Conclusions

Bolt-ons might represent a useful solution to improve the validity of a measure like the EQ-5D, while ensuring comparability between health technology assessments. However, studies conducted to date differed in the methods used to identify and select bolt-ons, and these methods had never been examined for this purpose. This cast doubts on the relevance of the bolt-on dimensions investigated and affected the possibility of comparing results from different studies. This thesis examined some of the possible methods that could be used to identify and select bolt-ons for GPBMs with the purpose of understanding what information they provide and how this can be used.

This research program found that literature reviews of psychometric evidence, commonly used for identifying bolt-ons, are often poor in quality, inconsistent in the tests reported and mostly focused on the EQ-5D. This affects the possibility of drawing comparative conclusions. Moreover, literature reviews of psychometric evidence generally report results only for the overall index of the measures examined. This undermines the possibility of identifying bolt-ons from this type of evidence. An alternative to identify bolt-ons is represented by the use of factor analysis. This thesis examined 4 techniques of the factor analysis family and found that these can be used to identify bolt-on dimensions for GPBMs. Having found some differences in the bolt-ons identified using PCA, EFA and CFA, the thesis suggested using them in conjunction. The thesis also found that SEM can be used to better understand the relationship between factors, but that this technique should be used with care and only in presence of a strong conceptual model, as different models might report similar goodness of fit indexes.

As the process of developing and appending bolt-ons might be costly and time consuming, this thesis investigated the possibility of reducing the list of available bolt-ons using the ability of bolt-ons to predict differences in HRQoL and the ability of bolt-ons to impact preferences for pairs of health states. Both techniques appeared useful to discriminate between candidate bolt-ons. However, there were differences in the bolt-ons suggested by these two methods, and these should be further explored.

This thesis represents the first effort ever conducted to compare methods for the purpose of bolt-ons identification and selection and it has contributed to the methodological knowledge of the bolt-on research area.

References

Aballea S, Tsuchiya A. Seeing for yourself: feasibility study towards valuing visual impairment using simulation spectacles. *Health Economics.* 2007.16:537-543

Aldred J. The sceptical economist. *Earthscan*. 2009.

American Hearth Association. Warning signs of heart failure. 2017. Available at: http://www.heart.org/HEARTORG/Conditions/HeartFailure/WarningSignsforHeartFailure/Warning-Signs-of-Heart-Failure_UCM_002045_Article.jsp#.WYHB5OmFw28

Allport GW. Personality: normal and abnormal. in Allport GW. Personality and Social Encounter. *Boston: Beacon Press.* 1964.

Arah OA, Westert GP, Hurst J et al. A conceptual framework for the OECD Health Care Quality Indicators Project. *International Journal of Quality of Health Care*. 2006. 18: 5–13.

Arnesen, T, Trommald, M. Are QALYs based on time trade-off comparable? – A systematic review of TTO methodologies. *Health Economics*. 2005. 14: 39–53.

Arons AMM, Krabbe PFM, van der Wilt GJ, Olde Rikkert GM, Adang EMM. Visual Analogue Scale: scale recalibration by patients with dementia and their proxies. *Quality of Life Research*. 2013. 22; 5: 979-86.

Arrindell WA, van der Ende J. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement.* 1985. 9: 165-178.

Arrow KJ. Uncertainty and the welfare economics of medical care. *American Economic Review*. 1963. 53: 941-967.

Ashby J, O'Hanlon M, Buxton MJ. The time trade-off technique: how do the valuations of breast cancer patients compare to those of other groups? *Quality of Life Research*. 1994. 3: 257–65.

Attema AE, Edelaar-Peeters Y, Versteegh MM, Stolck EA. Time trade off: One methodology, different methods. *European Journal of Health Economics*. 2013. 14; 1: S53-S64.

Bakker C, Rutten M, van Doorslaer E, Bennett K, van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *Journal of Rheumatology*. 1994. 21: 269–74.

Bansback N, Ara R, Karnon J, Anis A. Economic evaluations in rheumatoid arthritis: a critical review of measures used to define health States. *Pharmacoeconomics.* 2008. 26; 5: 395-408.

Bansback N, Brazier J, Davies S. Using contrast sensitivity to estimate the cost effectiveness of verteporfin in patients with predominantly classic age related macular degeneration. *Eye*. 2007. 21; 12: 1455-63.

Bansback N, Hole AR, Mulhern B, Tsuchiya A. Testing a discrete choice experiment including duration to value health states for large descriptive systems: Addressing design and sampling issues. *Social Science and Medicine*. 2014. 114: 38-48.

Bansback N, Tsuchiya A, Brazier J, Anis A. Canadian valuation of EQ-5D health states: preliminary value set and considerations for future valuation studies. *PLoS One*. 2012. 7.

Bartlett MS. Tests of significance in factor analysis. *British Journal of Psychology*. 1950. 3: 77-85.

Baron J, Greene J. Determinants of insensitivity to quantity on valuation of public goods: contribution, warm glow, budget constraints, availability and prominence. *Journal of Expert Psychological Applications*. 1996. 2: 106–125.

Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Comparing the performance of the EQ-5D-3L and SF-6D when measuring the benefits of alleviating knee pain. *Cost Effective Resource Allocation*. 2009. 7: 12.

Bateman I, Munro A, Rhodes B, Starmer C, Sugden R. Does part-whole bias exist? An experimental investigation. *The Economic Journal*. 1997. 107: 322–33.

Beattie J, Covey J, Dolan P, Hopkins L, Jones-Lee M, Loomes G, Pidgeon N, Robinson A, Spencer A. On the contingent valuation of safety and the safety of contingent valuation: part 1 caveat investigator. *Journal of Risk and Uncertainty*. 1998. 17: 5–26.

Beauducel A, Herzberg PY. On the performance of maximum likelihood versus variance adjusted weighted least squares estimation in CFA. *Structural Equation Modelling: A Multidisciplinary Journal*. 2006. 13: 186–203.

Bentler PM, Dudgeon P. Covariance structure analysis: statistical practise, theory and directions. *Annual review of psychology*. 1996. 47: 541-70.

Bentler PM, Kano Y. On the equivalence of factors and components. *Multivariate Behavioral Research*. 1990. 25: 67-74.

Bice T. Comments on health indicators. A methodological perspective. *International Journal of Health Services*. 1976. 6; 3: 509-20.

Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *Journal of Health Economics.* 1997. 16: 155–75.

Böckerman P, Johansson E, Saarni SI. Do established health-related quality-of-life measures adequately capture the impact of chronic conditions on subjective well-being?. *Health Policy*. 2011. 100: 91-5.

Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technology Assessessment*. 2014. 18; 34: 1-188.

Brazier, J, Deverill, M, A checklist for judging preference-based measures of health related quality of life: Learning from psychometrics. *Health Economics*. 1999. 8: 41-51.

Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. NHS R&D HTA programme: *Health Technology Assessment*. 1999. 3; 9.

Brazier J, Dolan P. Evidence of preference construction in a comparison of variants of the standard gamble method. *HEDS discussion paper*. 2005. 05/04.

Brazier J, Longworth L. NICE DSU Technical Support Document 8: An introduction to the measurement and valuation of health for NICE submission. *Report by Decision Support unit.* 2011. Available at: http://www.nicedsu.org.uk/TSD8%20Introduction%20to%20MVH_final.pdf

Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation. *Oxford University Press 1st ed.* 2007.

Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health for economic evaluations. *Oxford University Press 2nd ed.* 2017.

Brazier JE, Roberts J. Estimating a preference based index from the SF-12. *Medical Care*. 2004. 42: 851-9.

Brazier J, Roberts J, Deverill M. The estimation of preference based single index measure for health from SF-36. *Journal of Health Economics*. 2002. 21: 271-92.

Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health economics*. 2004.

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.

Brazier J, Rowen D, Tsuchiya A, Yang Y, Young TY. The impact of adding an extra dimension to a preference-based measure. *Social science and medicine*. 2011. 73: 245-53.

Brazier, J, D. Rowen Y L, Yang Y, Tsuchiya A. Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale. *European Journal of Health Economics. 2012.* 13; 5: 575-587.

Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology*. 1998. 51: 1115-18.

Brazier J, Tsuchiya A. Preference-based condition specific measures of health: what happens to cross programme comparability. *Health Economics*. 2010. 19; 2: 125-9.

Brod M, Tesler L, Christensen T. Qualitative research and content validity: developing best practices based on science and experience. *Quality of Life Research.* 2009; 18: 1263–78.

Brouwer WBF, Culyer AJ, Van Exel NJA, Rutten FFH. Welfarism versus extra-welfarism. *Journal of Health Economics*. 2008. 27: 325-38.

Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS. Testing structural equation models. *SAGE*. 1993. 136–62.

Buchanan J, Wordsworth S. Welfarism versus Extra-welfarism: Can the choice of economic evaluation approach impact on the adoption decision recommended by economic evaluation studies?. *Pharmacoeconomics*. 2015. 33: 571-79.

Buchanan JM, Tullock G. The Calculus of Consent. *Ann Arbor Paperback. Liberty Fund. Michigan.* 1962.

Buckingham JK, Birdsall J, Douglas JG. Comparing three versions of the time trade-off: time for a change? *Medical Decision Making.* 1996.

Busino G. Pareto, Vilfredo. *The new Palgrave dictionary of Economics*. London: Palgrave. 1987. 3: 799-804.

Cadman D, Goldsmith C, Torrance G, Boyle M, Furlong W. Development of health status index for Ontario children: final report to the ministry of health. *Grant research DM648. Mcmaster University, Hamilton, Ontario.* 1986.

CADTH database search filters. Canadian Agency for Drugs and Technologies in Health; Ottawa. 2014. 17/02/2015. Available from:: http://www.cadth.ca/en/resources/finding-evidence-is/string

Carson RT. Constructed markets. In: Braden JB, Kolstad CD. Measuring the demand for environmental quality. *Amsterdam: North-Holland*. 1991. 121–62.

Castelino M, Abbott J, McElhone K, The LS. Comparison of the psychometric properties of health-related quality of life measures used in adults with systemic lupus erythematosus: a review of the literature. *Rheumatology*. 2013. 52: 684.

Ching S, Thoma A, McCabe RE, Antony MM. Measuring outcomes in aesthetic surgery: a comprehensive review of the literature. *Plastic Reconstruction Surgery.* 2003. 111; 1: 469-80.

Clausen JA. Values, norms, and the health called 'mental': purposes and feasibility of assessment. in Sells SB. Definition and Measurement of Mental Health. *Washington, D.C.: U. S. Department of Health, Education and Welfare, National Center for Health Statistics. 1968.*

Cohen J. Statistical power analysis for the behavioural sciences. *New York: Academic Press.* 1977.

Comrey AL, Lee HB. A first course in factor analysis. *Hillsdale, NJ: Lawrence Erlbaum Associates 2nd edition*. 1992.

Cookson R. Willingness to pay methods in health care: a sceptical view. *Health economics.* 2003. 12; 11: 891–4.

Costelo AB, Osborne JW. Best practises in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research and evaluation*. 2005. 10; 7.

Culyer AJ. Introduction. In: Towse A, Pritchard C, Devlin N. Cost effectiveness thresholds: economic and ethical issues. *London: King's Fund and Office of Health Economics.* 2002: 9-15.

Culyer AJ. Introduction. In Health Indicators: An International Study for the European Science Foundation. *New York: St. Martin's.* 1983.

Culyer, AJ. The normative economics of health care finance and provision. In: McGuire, A., Fenn, P., Mayhew, K. (ed), Providing Health Care: The Economics of Alternative Systems of Finance and Delivery. *Oxford University Press: Oxford.* 1991.

D'Agostino RB, Russell HK. Scree test. *Encyclopedia of Biostatistics.* 2005.

Davis S, Wailoo A. A review of the psychometric performance of the EQ-5D in people with urinary incontinence. *Health Quality of Life Outcomes.* 2013. 18; 11.

Delquie P. 'Bi-matching'': a new preference assessment method to reduce compatibility effects. *Management Science.* 1997. 43: 640–58.

Derrett S, Black J, Herbison GP. Outcome after injury-a systematic literature search of studies using the EQ-5D. *Journal of Trauma*. 2009. 67; 4: 883-90.

Detsky AS, McLaughlin JR, Abrams HB, L'Abbe KA, Whitwell J, Bombardier C, et al. Quality of life of patients on long-term total parenteral nutrition at home. *Journal of General Internal Medicine*. 1986. 1: 26–33.

DeVine J, Norvell DC, Ecker E, Fourney DR, Vaccaro A, Wang J, et al. Evaluating the correlation and responsiveness of patient-reported pain with function and quality-of-life outcomes after spine surgery. *Spine (Phila Pa 1976).* 2011. 36; 21: S69-74.

Devlin NJ, Hansen P, Selai C. Understanding health state valuations: a qualitative analysis of respondents' comments. *Quality of Life Research.* 2004. 13; 7: 1265-77.

Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health related quality of life: A value set for England. *Office of Health Economics Research Paper 16/01*. 2016. Available at: file:///Users/Aureliano/Downloads/25%20Jan%20%20OHE%20research%20paper_value%20set%20paper%20CORRECTED.pdf

Devlin, N., Tsuchiya, A., Buckingham, K. and Tilling, C. A uniform time trade off method for states better and worse than dead: Feasibility study of the ‘lead time’ approach. *Health Economics.* 2011. 20; 3: 348-61.

Diener E, Suh EM, Lucas RE, Smith HL. Subjective wellbeing. Three decades of progress. *Pshycological Bullentin.*1999. 125; 2: 276-302.

Dolan P. Developing methods that really do value the ‘‘Q’’ in the QALY. *Health Economics Policy and Law.* 2008 (a). 3: 69–77.

Dolan P. In defence of subjective well-being. *Health Economics Policy and Law.* 2008 (b). 3: 93–5.

Dolan P, Kahnemann D. Interpretations of utility and their implications for the valuation of health. *The Economic Journal.* 2008. 525: 215-34.

Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Economics.* 1995. 4: 289-99.

Dolan P, Gudex C, Kind P, Williams A. Valuing Health States: a comparison of methods. *Journal of health economics.* 1996. 15: 209-31.Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and people's preferences: A methodological review of the literature. *Health Economics. 2005.* 2: 197-208.

Donaldson C, Gerard K. Countering moral hazard in public and private health care systems: a review of recent evidence. *Journal of Social Policy.* 1989. 18: 235-51.

Dowie J. Decision validity should determine whether generic or condition specific HRQOL measures is used in health care decisions. *Health economics.* 2002. 11: 1-8.

Drummond MF. Output measurement for resource allocation decisions in health. *Oxford Review of Economic Policy.* 1989. 5: 59-74.

Drummond M, McGuire A. Economic Evaluation in Health Care: Merging Theory with Practice. *Oxford University Press: Oxford*. 2001.

Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes, *Oxford University Press: Oxford 2nd ed.* 2005.

Drummond MF, Stoddard GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programme. *Oxford: Oxford Medical Publications 1st ed.* 1987.

Dubos R. Man, medicine and environment. *Harmondsworth: Pelican.* 1968.

Dunn HL. High Level Wellness, *Beatty Press. Arlington, VA.* 1961.

Dyer MTD, Goldsmith KA, Sharples LS, Buxton MJ. A review of health utilities using the EQ5D in studies of cardiovascular disease. *Health and Quality of Life Outcomes.* 2010.

Edwards S., Lilford, R. and Kiauka, S. Different types of systematic review in health services research. In N. Black et al. eds. *Health Services Research Methods: a Guide to Best Practice. London, BMJ Books*. 1998. 255-59.

Essink-Bot ML, Stuifbergen MC, Meerding WJ, Looman CW, Bonsel GJ. Individual differences in the use of the response scale determine valuations of hypothetical health states: an empirical study. *BMC Health Services Research.* 2007. 7; 62.

EuroQol group. EuroQol – A new facility for the measurement of health related quality of life. *Health policy.* 1990. 16: 199-208.

EuroQol Group. User Guide. Basic information on the use of the EQ-5D. *EuroQol Group*. 2009.

EuroQol Group symposium. New research for future approaches to measuring and valuing health. *International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Berlin, Germany*. 2012.

Evans RG. Strained mercy: The economics of Canadian health care. *Toronto: Butterworths.* 1984.

Evans RG. Supplier-induced demand: Some empirical evidence and implications, in: Perlman M. The economics of health and medical care. *New York: John Wiley and Sons.* 1974. 162-73.

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*. 1999. 4: 272.

Faulkner EJ. Health Insurance. *New York.* 1960. 327.

Farivar SS, Cunnigham WE, Hays RD. Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey. *Health and Quality of Life Outcomes*. 2007. 5: 54.

Fava G, Ruini C. Development and characteristics of a well-being enhancing psychotherapeutic strategy: Well-being therapy. *Journal of Behavioural Therapy and Experimental Psychology.* 2003. 34: 45-63.

Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient reported outcomes. *John Wiley and Sons.* 2007.

Feldman R, Dowd B. A new estimate of the welfare loss of excess health insurance. *American Economic Review.* 1991. 81: 297–301.

Feldman R, Sloan F. Competition among physicians, revisited. In: Greenberg, W. Competition in the Health Care Sector: Ten Years Later. *Duke Univ. Press.* 1988.

Feldstein M. The welfare loss of excess health insurance. *Journal of Political Economy.* 1973. 81: 251–80.

Feng Y, Parkin D, Devlin NJ. Assessing the performance of the EQ-VAS in the NHS PROMs program. *Quality of Life Research.* 2014. 23: 977-989.

Feeny D. The utility approach to assessing population health. In Murray C, Salomon J, Mathers C, Poez A, Lozano R. Sumary measures of population health: concepts, ethics, measurements and applications. *World Health Organization: Geneva.* 2002. 515-28.

Feeny DH, Furlong WJ, Torrance GW, Goldsmith CH, Zenglong Z Depau S, Denton M, Boyle M. Multi-attribute and single attribute utility function the health utility index mark 3 system. *Medical Care.* 2002. 40: 113-28.

Field A. Discovering statistics using SPSS. *Sage 3rd ed.* 2009.

Finch AP, Brazier JE, Mukuria C, Bjorner J. An exploratory study on using principal component analysis and confirmatory factor analysis to identify bolt-on dimensions: The EQ-5D case study. *Value in Health.* Jun 2017 (a). (In print) .

Finch AP, Brazier J, Mukuria C. What is the evidence for the performance of Generic Preference-Based Measures? A systematic overview of reviews. *European Journal of Health Economics. 2017 (b)* (In print).

Folland S, Goodman, AC, Stano, M, The Economics of Health and Health Care, *Prentice Hall: New Jersey. 5th ed.* 2007.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – II: scaling methods. *Journal of Clinical Epidemiology*. 1989. 42: 459–71.

Fuchs V. The supply of surgeons and the demand for operations. *Journal of Human Resources*. 1978. 13: 35-56.

Fuchs VR. Physician-induced demand: A parable. *Journal of Health Economics.* 1986. 5: 367.

Furlong W, Feeny D, Torrance G, Barr R, Horsman J. Guide to design and development of health-state utility instrumentation. 1992.

Gafni, A. Economic Evaluation of Health-care Programmes: Is CEA Better than CBA? *Environmental & Resource Economics*. 2006. 34; 3: 407–18.

Gafni A. Willingness to pay as a measure of benefits: relevant questions in the context of public decision making about healthcare programs. *Medical Care.* 1991. 29; 12: 1246-52.

Gafni A, Charles C, Whelan T. The physician patient encounter: they physician as a perfect agent for the patient versus the informed treatment decision making model. *Social Science and Medicine.* 1998. 47; 3: 347-54.

Garner L. The NHS: Your Money or Your Life. *Penguin: New York.* 1979.

Garrison Jr, LP. Editorial: On the benefits of modelling using QALYs for societal resource allocation. The model is the message. *Value in Health.* 2009. 12: S36-37.

Glik HA, Doshi JA, Sonnad SS, Polski D. Economic evaluations in clinical trials. *Oxford University Press, Oxford*. 2005.

Gold MR, Patrick DL, Torrance GW, Fryback D, Hadorn DC, Kamlet M, et al. Identifying and valuing outcomes. In: Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in health and medicine. *Oxford University Press: Oxford*. 1996.

Gorsuch RL. Common factor analysis versus principal component analysis: Some well and little known facts. *Multivariate Behavioral Research*. 1990. 25: 33-39.

Graham C. Happiness and health: lessons and questions for public policy. *Health Affairs.* 2008. 27; 1: 72-87.

Grbich C. Qualitative research in health: An introduction. *St Leonards, N.S.W.: Allen & Unwin.* 1999.

Greer A. The measurement of health in urban communities. *Journal of Urban Affairs.* 1986. 8; 11.

Green C, Brazier J, Deverill M. A review of health state valuation techniques. *Pharmacoeconomics*. 2000. 17: 151-65.

Gudex C. Are we lacking a dimension of energy in the EuroQol instrument? *Paper presented at the 8th Pleanary Meeting of the EuroQol Group, Lund, Sweden.* 1991.

Gudex C, Dolan P, Kind P, et al. Health state valuations from the general public using the visual analogue scale. *Quality of Life Research.* 1996; 5: 521-31.

Hair J, Anderson RE, Kaiser HF. A Second-Generation Little Jiffy. *Psychometrika*. 1970. 35; 4: 401-15.

Haunsman DM, McPherson MS. Taking ethics seriously: Economics and Contemporary Moral Philosopy. *Journal of Economic Literature*. 1993. 31; 2: 671-731.

Haute Autorite de Sante. Choices in methods for economic evaluation. A methodological guide. 2012. Available at: http://www.has-sante.fr/portail/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf

Haybron DM. The pursuit of unhappiness: the elusive psychology of well-being. *Oxford University Press: Oxford*. 2008.

Hays RD, Morales LS. The RAND-36 measure of health-related quality of life. *Annals of Medicine.* 2001. 33: 350–357.

Hays RD, Prince-Embury S, Chen H. R-36 HIS: RAND-36 Health Status Inventory. *The Psychological Corporation: San Antonio, TX*. 1998.

Hays RD, Reeve BB. Measurement and modeling of health-related quality of life. In: Killewo J, Heggenhougen HK, Quah SR. Epidemiology and demography in public health. *Academic Press: San Diego*. 2010. 195–205.

Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. *Health Economics*. 1993. 2: 217–27.

Haywood KL, Garratt AM, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. *Quality of Life Research.* 2005. 14; 7: 1651-68.

Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. *Quality of Life Research.* 1999. 8; 3: 209–24.

Heintz E, Krol M, Levin LA. The impact of patients' subjective life expectancy on time trade-off valuations. *Medical Decision Making.* 2013. 33: 261–70.

Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five level version of EQ-5D (EQ-5D-5L). *Quality of Life Research.* 2011. 20; 10: 1727-36.

Hicks JR. The foundations of welfare economics. *Economic Journal.* 1939. 49: 696–712.

Hicks, J. R. The rehabilitation of consumers' surplus. *Review of Economic Studies.* 1941. 8: 108–16.

Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. *The Cochrane Collaboration. 2011.* Available from www.cochrane-handbook.org.

Hill MR, Noonan VK, Sakakibara BM, Miller WC, SCIRE Research Team. Quality of life instruments and definitions in individuals with spinal cord injury: a systematic review. *Spinal Cord.* 2010. 48; 6: 438-450.

Hodgson, GM. Towards an alternative economics of health care. *Health Economics, Policy and Law.* 2009. 4: 99-114.

Holloway L, Humphrey L, Heron L, Pilling C, Kitchen H, Hojbjerre L, et al. Patient-reported outcome measures for systemic lupus erythematosus clinical trials: a review of content validity, face validity and psychometric performance. *Health and Quality of Life Outcomes.* 2014. 12.

Hoogendoorn M, Oppe M, Boland M, Goossens S, Stolk E, Rutten Van Molken M. Exploring the impact of adding a respiratory dimension to the EQ-5D-5L. *Paper presented at the 33rd EuroQol Plenary meeting, Berlin, Germany.* 2016.

Hounsome N, Orrell M, Edwards RT. EQ-5D as a quality of life measure in people with dementia and their carers: evidence and key issues. *Value Health.* 2011; 14: 2: 390-99.

Hu LT, Bentler PM: Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model 1999, 6: 1–55.

Hurley J. An overview of the normative economics of the health care sector. In: Culyer AJ, Newhouse J. Handbook of Health Economics. *Elsevier Science: Amsterdam.* 2000.

Hurley, J. Welfarism, extra-welfarism and evaluative economic analysis in the health care sector. In: Barer ML, Getzen TE, Stoddard GL. Health, Health Care and Health Economics: Perspectives on Distribution. *John Wiley & Sons Ltd: Chichester.* 1998.

Institute of Medicine. Crossing the Quality Chasm: A new Health System for the 21st Century. *Washington, DC: National Academy Press.* 2001.

Jahoda M. Current Concepts of Positive Mental Health. *Basic Books: New York.* 1958.

James C, Carrin C, Savedoff W, Hanvoravongchai P. Clarifying Efficiency-Equity trade-offs through explicit criteria, with a focus on developing countries. *Health Care Analysis.* 2005. 13: 1.

James M, Stokes E. Harnessing Information for Health Economics Analysis. *Radcliffe Oxford.* 2006.

Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with Type 2 diabetes mellitus. *Diabetic Medicine.* 2011. 28; 4: 395-413.

Jelsma J, Maart S. Should additional domains be added to the EQ-5D health related quality of life instrument for community based studies? An analytical descriptive study. *Population Health Metrics.* 2015. 13; 13: 1-11.

Johannesson M. QALYs, HYEs and individual preferences – a graphical illustration. *Social Science and Medicine.* 1994. 39: 1623–32.

Johannesson M. Theory and methods of economic evaluation of health care. *Dordrecht: Kluwer*. 1996.

Johannesson M, Jonsson B. Economic evaluation in health care: is there a role for cost-benefit analysis? *Health Policy.* 1991. 17: 1–23.

Johannesson M, Jönsson B, Karlson G. Outcome measurement in economic evaluation. *Health Economics.* 1996. 5: 279–96.

Johnson FR. Why not real economics? *PharmacoEconomics.* 2012. 30; 2: 127–31.

Jones-Lee M, Loomes G, O'Reilly D, Phillips P. The value of preventing non-fatal road injuries: findings of a willingness to pay national sample survey. *Transport Research Laboratories*. 1993.

Kahneman D, Knetsch JL. Valuing public goods: the purchase of moral satisfaction. *Journal of Environmental Economical Management.* 1992. 22: 57–70.

Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979; 47: 263–91.

Kaiser HF. The application of electronic computers to factor analysis. *Educational and Psychological Measurement.* 1960. 20: 141-51.

Kaiser HF, Rice J. Little jiffy. Mark 4. *Educational and Psychological Measurement.* 1974. 34: 111-7.

Kalarman H. Application of cost-benefit analysis to the health services and the special case of technology innovation. *International Journal of Health Services.* 1974. 4: 325.

Kaldor N. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal.* 1939. 49: 549–52.

Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Quality of Life Research.* 1993. 2: 467-75.

Karimi M, Brazier JE. Health, Health related quality of life and quality of life: What is the difference? *Pharmacoeconomics*. 2016. 34; 7: 645-49.

Karimi M, Brazier JE, Paisley S. Are preferences over health states informed? *Health and Quality of Life Outcomes.* 2017.

Kelman HR, Willner A. Problems in measurement and evalua- tion of rehabilitation. *Archives of Physical Medicine and Rehabilitation*. 1962. 43: 172-181.

Kline RB. Principles and Practise of Structural Equation Modelling. *The Guilford Press: New York*. 2016.

Konerding U, Moock J, Kohlmann T. The classification system of the EQ-5D, HUI II and SF-6D: what do they have in common?. *Quality of life research*. 2009. 18: 1249-61.

Hausman, JA, Mc Pherson (ed.). In Kemp MA, Maxwell C. Exploring a budget context for contingent valuation estimates. In Contingent Valuation: a Critical Assessment. Chapter 5. *Cambridge Economics Inc: Cambridge MA*. 1993. 217–65.

Kharroubi SA, O'Hagan A, Brazier EJ. Estimating utilities from individual health preference data: a non parametric Bayesian method. *Applied Statistics*. 2005. 54: 879-95.

Killewo J, Heggenhougen K, Quah SR. Epidemiology and demography in public health. *Academic Press: Amsterdam*. 2010.

Klose T. The contingent valuation method in health care. *Health Policy*. 1999. 47: 97-123.

Kouba P. Conceptualizing health and illnesses. *Journal of Phenomenological Psychology*. 2008. 39: 59-80.

Krabbe PFM, Stouthard MEA, Essink-Bot ML, Bonsel GJ. The Effect of Adding a Cognitive Dimension to the EuroQol Multiattribute Health-Status Classification System. *Journal of Clinical Epidemiology*. 1999. 52: 293-301.

Krucien N, Ryan M, Hermens F. Using eye-tracking methods to inform decision making processes in discrete choice experiments. *Health Economist's Study Group*. 2014.

Kuppermann M, Shiboski S, Feeny D, et al. Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application to prenatal diagnosis. *Medical Decision Making*. 1997. 17: 42-55.

Kuspinar A, Mayo NE. A review of the psychometric properties of generic utility measures in multiple sclerosis. *Pharmacoeconomics.* 2014. 32; 8: 759-73.

Larson JS. The conceptualization of health. *Medical care research and review*. 1998. 56; 2: 123-36.

Larson JS. The measurement of health. *Greenwood Publishing Group.* 1991.

Ledesma RD, Valero-Mora PV. Determining the number of factors to retain in EFA: an easy to use computer program for carrying out parallel analysis. *Pratical assessment, research and evaluation*. 2007. 12; 2.

LeGales C, Buron C, Costet N, Rosman S. Development of preference-weighted health status classification system in France: the Health Utility Index 3. *Health Care Management Science.* 2002. 5: 41-51.

Lerner M. Conceptualization of health and social well-being. *Spring*. 1973. 8; 1: 6-12.

Labelle R, Stoddart G, Rice T. A re-examination of the meaning and importance of supplier induced demand. *Journal of Health Economics.* 1994. 13: 347-68.

Lenert LA, Cher DJ, Goldstein MK, Bergen MR, Garber A. The effect of search procedures on utility elicitations. *Medical Decision Making.* 1998. 18: 76–83.

Lenert L, Sturley AP, Rapaport MH, Chavez S, Mohr PE, Rupnow M. Public preferences for health states with schizophrenia and a mapping function to estimate utilities from positive and negative scores. *Schizophrenia Research*. 2004. 71: 83-95.

Lewis A. Health as a social concept. *British Journal of Sociology. 1953*. 4: 109-124.

Linder JA, Singer DE, Ancker M, Atlas SJ. Measures of health-related quality of life for adults with acute sinusitis. A systematic review. *Journal of General Internal Medicine.* 2003, 18; 5: 390-401.

Lloyd A. Threats to the estimation of benefit: are preference estimation methods accurate?. *Health Economics*. 2003. 12: 393-402.

Lloyd A, Doyle S, Dewild S, Turk F. Preferences and utilities for the symptoms of moderate to severe allergic asthma. *European Journal of Health Economics.* 2008. 9: 275-84.

Longworth L, Mulhern B, Yang Y, et al. A systematic review of the performance of EQ-5D in four disease areas. *Poster presented at 28th Scientific Plenary Meeting of the EuroQol Group, Oxford, UK.* 2011.

Longworth L, Yang Y, Young T, Mulhern B, Hernandez Alava M, Mukuria C, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technology Assessment.* 2014. 18; 9: 1-224.

Loomes G, McKenzie L. The use of QALYs in health care decison making. *Social Science and Medicine.* 1989. 28: 299–308.

Louviere JJ, Hensher DA, Swait JD. Stated choice methods: analysis and applicaiton. Cambridge University Press. Cambridge. 2000.

Luo N, Wang X, Ang N, Finkelstein EA, Aung T, Wong T, Lamoureux E. A vision bolt-on item could increase the discriminatory power of the EQ-5D index score. *Value in Health.* 2015. 18: 1037-42.

Marin A, Psacharopoulis G. The reward for risk in the labour market: evidence from the UK and a reconciliation with other studies. *Journal of Political Economics.* 1982. 90: 827-853.

Marvin D, Crown RW. Concepts And Surveys For Area Health Care Planners. *Ames, Iowa: Center For Agricultural And Rural Development, Iowa State University.* 1976.

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. *Oxford University Press: Oxford.* 1987.

McGuire A, Henderson J, Mooney G. The economics of health care. *Routledge: New York.* 1988.

McPhail S, Haines T. Response shift, recall bias and their effect on measuring change in health related quality of life amongst older hospital patients. *Health and Quality of Life Outcomes.* 2010. 8; 65.

Mechanic D. Medical Sociology: A Selective View. 1968. *The Free Press: New York.*

Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Medical Decision Making*. 1994.13: 287–92.

Meiselman . Quality of life, well-being and wellness: measuring subjective health for foods and other products. *Food, quality and preferences.* 2016. 54:101-109.

Mishan E. Evaluation of life and limb: a theoretical approach. *Journal of Political Economy.* 1971. 79: 687 – 705.

Mitchell RC, Carson RT. Using Surveys to Value Public Goods: the Contingent Valuation Method. *Resources for the Future: Washington.* 1990.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine.* 2009. 6; 6. Available from: http://www.prisma-statement.org/2.1.2%20-%20PRISMA%202009%20Checklist.pdf

Mold JW. An alternative conceptualization of health and healthcare: its implication for geriatrics and gerontology. *Educational Gerontology*. 1995. 21; 1: 85-101.

Moock J, Kohlmann T. (2008). Comparing preference-based quality-of-life measures: results from rehabilitation patients with musculoskeletal, cardiovascular, or psychosomatic disorders. *Quality of Life Research.* 17: 485–95.

Mooney G. Economics, medicine and health care. *Health economics*. 2003.

Moons P, Budts W, De Geest S. Critique on the conceptualization of quality of life: A review and evaluation of different conceptual approaches. *International Journal of Nursing Studies*. 2006. 43: 891-901.

Morris S, Devlin N, Parkin D. Economic Analysis in Health Care. *John Wiley and Sons*. 2007.

Mulaik SA. Blurring the Distinctions between Component Analysis and Common Factor-Analysis. *Multivariate Behavioral Research*. 1990. 25; 1: 53-59.

Mulhern B, Bansback N, Brazier J, Buckingham K, Cairns J, Devlin N, Dolan P, Hole AR, Kavetsos G, Longworth L, Rowen D, Tsuchiya A. Preparatory study for the re-valuation of the EQ-5D tariff: Methodology Report. *Health Technology Assessment,* 2014. 18; 12.

Mulhern B, Longworth L, Brazier B, Rowen D, Bansback N, Devlin N, Tsuchiya A. Binary Choice Health State Valuation and Mode of Administration: Head-to-Head Comparison of Online and CAPI. Value Health. 2013 16; 1: 104–13.

Mulhern B, Norman R, Viney R, Stolk E. Using discrete choice experiments to value generic preference based measures: A systematic review. Value in health. 2016. 19. A95.

Murphy R, Sackley CM, Miller P, Harwood RH. Effect of experience of severe stroke on subjective valuations of quality of life after stroke. *Journal of Neurology, Neurosurgery and Psychiatry.* 2001. 70: 679-81.

Muthen BO, Muthen L. Mplus User's Guide. *Muthén and Muthén: Los Angeles*. 1998.

Nagi SZ. Disability and Rehabilitation. *Ohio State University Press: Columbus, Ohio*. 1969.

National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. *National Institute for Health and Care Excellence.* 2013. Available at: http://www.nice.org.uk/article/pmg9/chapter/Foreword

Navarro, V. Health and Medical Care in the U.S.: A Critical Analysis. *Farmingdale, NY: Baywood.* 1977.

Newhouse J, the Health Insurance Experiment Group. Free for All? Lessons from the RAND Health Insurance Experiment. *Harvard University Press: Cambridge, MA.* 1993.

Newhouse JP. The Structure of Health Insurance and the Erosion of Competition in the Medical Marketplace. In: Competition in the health care sector: Past, present, future. *Greenberg W. Germantown. Aspen Systems Corporation.* 1978. 215-30.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *International Journal of Health Planning and Management.* 1991; 6: 234–42.

Nord E. Methods for Quality Adjustment of Life Years. *Social Science and Medicine.* 1992. 94; 5: 559-69.

Nord E, Richardson J, Macarounas-Kirchmann K. Social evaluation of health care versus personal evaluation of health states. *International Journal of Technology Assessment in Healthcare.* 1993; 9: 463-468.

Norman R, King MT, Clarke D, et al. Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. *Quality of Life Research.* 2010. 19: 499– 508.

Nyman JA. The economics of moral hazard revisited. *Journal of Health Economics.* 1999. 18: 811-82.

O'Brien B, Gafni A. When do the dollars make sense? Towards a conceptual framework for contingent valuation studies. *Medical Decision Making.* 1996. 16: 288–99.

O'Brien B, Viramontes JL. Willingness to pay: a valid and reliable measure of health state preferences? *Medical Decision Making.* 1994. 14; 3: 289-97.

Office for National Statistics. Measuring Subjective Well-being. *Office for National Statistics.* 2011.

Olsen JA, Smith RD. Theory versus practice: a review of “willingness-to-pay” in health and health care. *Health Economics*. 2001. 10; 1: 39–52.

Papaioannou D, Brazier J, Parry G. How to measure quality of life for cost effectiveness analysis in personality disorder? A systematic review. *HEDS discussion paper 13/02.* 2013.

Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. *Value Health.* 2011. 14; 6: 907-20.

Parducci A. Contextual effects. A range-frequency analysis. In: Carterette E, Friedman M. Handbook of perception. vol 2. *Academic Press: New York.* 1974.

Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost utility analysis?. *Health Economics.* 2006. 15: 653-64.

Parliamentary Well-being group. 2014. Available from: http://parliamentarywell-beinggroup.org.uk

Parsons T. Definitions of health and illness in light of American values and social structure. in Gartley Jaco E. Patients, Physicians, and Illness. *New York: The Free Press. 1958.*

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Medical Decision Making.* 1994. 14: 9–18.

Paul-Dauphin A, Guillemin F, Virion JM, Briancon S. Bias and precision in visual analogue scales: A randomized controlled trial. *American Journal of Epidemiology.* 1999. 150; 10: 1117-27.

Pauly MV. Indemnity Insurance for Health Care Efficiency. *Economics and Business Bulletin.* 1971. 4: 53-9.

Pauly MV. Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection. *Quarterly Journal of Economics.* 1974. 88; 1: 44-62.

Pauly MV. Taxation, Health Insurance, and Market Failure in the Medical Economy. *Journal of Economic Literature.* 1986. 24; 2: 629-75.

Pauly MV. The economics of moral hazard. *American Economic Review.* 1968. 58; 3: 531-37.

Payne L. Health. A basic concept in nursing theory. *Journal of Advancing Nursing.* 1983. 8: 393-95.

Peasgood T, Brazier J, Papaioannou D. A systematic review of the validity and responsiveness of EQ5D and SF6D for depression and anxiety. *HEDS discussion paper 12/15 (unpublished).* 2012.

Peasgood T, Ward SE, Brazier J. Health state utility values in brest cancer. *Expert review of Pharmacoeconomics and Outcome Research.* 2010. 10: 553-66.

Pedhazur EJ, Schmelkin LP. Measurement, Desing and Analysis: An integrated approach. *Lawrence Erlabaum Associates:* New Jersey. 1991.

Perneger TV, Courvoisier DS. Exploration of health dimensions to be included in multi-attribute health utility assessment. *International Journal for Quality in Healthcare.* 2011. 23; 1: 53-9.

Petrillo J, van Nooten F, Jones P, Rutten-van Molken M. Utility estimation in chronic obstructive pulmonary disease: a preference for change? *PharmacoEconomics.* 2011. 29; 11: 917-32. Epub 2011/10/13.

Pfaff M, Wassener, D. Germany. In: Rice TH, Biles B, Brown ER, Diderichsen F, Kuehn H. Reconsidering the role of competition in health care markets: an introduction. *Journal of Health Politics, Policy and Law.* 2000. 25; 5: 863-73.

Phelps CE. Health economics. *Harper Collins: New York*. 1992.

Pickard AS, De Leon MC, Kohlmann T et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care.* 2007. 45: 259–263.

Pickard AS, Wilke C, Jung E, Patel S, Stavem K, Lee TA. Use of a preference-based measure of health (EQ-5D) in COPD and asthma. Respiratory Medicine. 2008. 102. 4: 519-536.

Pickard AS, Wilke CT, Lin HW, Lloyd A. Health utilities using the EQ-5D in studies of cancer. *Pharmacoeconomics*. 2007. 25; 5: 365-384.

Plucker JA. Exploratory and confirmatory factor analysis in gifted education: examples with self concept data. *Journal of the education of the gifted.* 2003. 27; 1: 20-35.

Post PN, Stiggelbout AM,Wakker PP. The utility of health states after stroke: a systematic review of the literature. *Stroke.* 2001; 32: 1425-9.

Redisch, M, Gabel J, Blaxall M. Physician pricing, costs, and income. In: Scheffler RM. Advances in health economics and health services research. *JAI Press: Greenwich.* 1981. 197-228.

Reinhardt, U. The theory of physician-induced demand. Reflections after a decade. *Journal of Health Economics.* 1985. 4: 187–93.

Revicki DA. Relationship between health utility and psychometric health status measures. *Medical Care.* 1992. 30: MS274–82.

Revicki DA, Wood M. Patient assigned health state utilities for depression related outcomes: differences by depression severity and antidepressant medication. *Journal of Affective Desorders.* 1998. 48: 25-36.

Rice TH, Labelle RJ. Do physicians induce demand for medical services? Journal of *Health Economics, Policy and Law.* 1989. 14: 587-600.

Richardson J. Cost utility analysis: what should be measured? *Social Science and Medicine.* 1994. 39: 7–21.

Richardson J, Chen G, Khan MA, Iezzi A. Can Multi-attribute Utility Instruments Adequately Account for Subjective Well-being? *Medical Decision Making.* 2015 (a).

Richardson J, Cummins R, Olsen J, Kaplan R, Coast J, Schlander M . A cross national comparison of 8 generic quality of life instruments. 2014 (a). Available online at: http://www.aqol.com.au/index.php/aqol-current

Richardson J, Iezzi A, Khan MA. Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and "micro-utility effects". *Quality of life research.* 2015 (b).

Richardson J, Iezzi A, Maxwell A. Cross-National Comparison of Twelve Quality of Life Instruments: MIC paper 1 Background, Questions, Instruments. *Centre for Health economics*. *Monash University.* 2012 (a).

Richardson J, Khan MA, Iezzi A, Maxwell A. Cross national comparison of twelve quality of life instruments. Research papers 78, 80-83, 85. MIC Report 2. *Centre for Health Economics, Monash University.* 2012 (b).

Richardson J, Khan M, Iezzi A, Maxwell A. Comparing and Explaining Differences in the Magnitude, Content, and Sensitivity of Utilities Predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D Multiattribute Utility Instruments. *Medical Decision Making.* 2014 (b).

Richardson J, Sinha K, Iezzi A, Munir KA. Modelling the utility of health states with the Assessment of Quality of Life (AQoL) 8 D instrument: overview and utility scoring algorithm. *Centre for Health Economics*. *Monash University.* 2011.

Robinson R. Cost-benefit analysis. *British Medical Journal.* 1993. 307: 924-26.

Robinson A, Covey J, Jones-Lee M, Loomes G. Comparing the Results of Face-to-face and Web Based PTO Exercises. *East Anglia, UK: Health Economics Study Group.* 2008.

Rossiter LF, Wilensky GR. A re-examination of the use of physician services: The role of physician-initiated demand. *Inquiry*. 1983. 20: 162-72.

Rotschild M, Stiglitz J. Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *Quarterly Journal of Economics.* 1976. 90: 630- 49.

Rowen D, Brazier J, van Hout B. A comparison of methods for converting DCE values onto the full health– dead QALY scale. *HEDS Discussion Paper*11/15. 2011.

Ruesch, J, Brodsky CM. The concept of social disability. *Archives of General Psychiatry*. 1968. 19: 394-404.

Ryan RM, Daci EL. To be happy or to be self-fulfilled: A review of researchon hedonic and eudaimonic well-being. In Fiske S. *Annual Review of Psychology: Palo Alto, CA*. 2001. 52: 141-66.

Ryff DC. Psychological well-being in adult life. Current directions in Psychological Science. 1995. 4: 99-104.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *Journal of Chronic Diseases*. 1981. 31: 697-704.

Safilios-Rothschild C. The Sociology and Social Psychology of Disability and Rehabilitation. *Random House: New York*. 1970.

Saha K, Mulhern B, Longworth L, Janssen B. Important aspects of health not captured by the EQ-5D. Views of the general public. *EuroQol working group series.* 2016.

Sanghera S, Frew E, Kai J, Gupta J, Elizabeth Roberts T. An assessment of economic measures used in menorrhagia: a systematic review. *Social Science and Medicine.* 2013. 98: 149-53.

Schulper M. Economic evaluation of minimal access surgery: the case of surgical treatment for menorrhagia. *HERG PhD thesis. Brunel University.* 1996.

Scott A. Giving things up to have more of others: the implications of limited substitutability for eliciting preferences in health and health care. *UK Health Economists Study Group Meeting.* 1998. Sheffield.

Sen AK. Social choice theory. In: Arrow KJ, Intriligator MD. Handbook of Mathematical Economics. vol. III. *Elsevier Science: Amsterdam.* 1986.

Shea BJ, Grimshaw, JM, Wells GA, Boers M, Anderson N et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology.* 2007; 7: 10. Available form: http://amstar.ca/Amstar_Checklist.php

Sintonen H. The 15D-Measure of Health-Related Quality of Life, II: Feasibility, Reliability and Validity of Its Valuation System. *National Centre for Health Program Evaluation: Melbourne, Australia*. 1995.

Slovic, P. The construction of preferences, *American Psychologist.* 1995. 50; 95: 364-71.

Smith MB. Social Psychology and Human Values. *Aldine Publishing: Chicago.* 1969.

Smith R. The end of disease and the beginning of health. *BMJ Group Blogs.* 2008. http:/ /blogs.bmj.com/bmj/2008/07/08/richard-smith-the-end-of-disease-and-the-beginning-of- health/.

Smith RD, Sach RD. Contingent valuation: (still) on the road of nowhere?. *Health Economics.* 2009. 18: 863-66.

Speight J, Reaney MD, Barnard KD. Not all roads lead to Rome-a review of quality of life measurement in adults with diabetes. *Diabetic Medicine.* 2009. 26; 4: 315-327.

Sprangers MA, Schwartz CE. The challenge of response shift for quality of life based clinical oncology research. *Annals of Oncology.* 1999. 10: 747-9.

Stano M. An analysis of the evidence on competition in the physician services markets. *Journal of Health Economics.* 1985. 4: 197-211.

Steiger H. Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika*. 1974. 14: 157-167.

Steiger JH. Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research*. 1990. 25: 173–80.

Stevens K, Palfreyman S. The use of qualitative methods in developing the descriptive system of preference based measures of health related quality of life for economic evaluation. *Value in Health*. 2012. 15; 8: 991-8.

Stiggelbout AM, Kiebert GM, Kievit, J, Leer, JW, Habbema JD, De Haes JC. The ‘‘utility’’ of the time trade-off method in cancer patients: feasibility and proportional trade-off. *Journal of Clinical Epidemiology.* 1995. 48: 1207–14.

Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making*. 1994. 14: 82–90.

Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. *Value in Health.* 2010. 13; 8: 1005-1013.

Stutzer A, Frey BS. Does marriage make people happy, or do happy people get married? *Journal of Socioeconomics*. 2003. 35; 2: 326-47.

Sutherland HJ, Llewellyn Thomas H, Boyd D, Till JE. Attitudes towards quality of survival: the concept of maximum endurable time. *Medical Decision Making*. 1982. 2: 299–309.

Swinburn P, Lloyd A, Boye KS, Edson-Heredia E, Bowman L, Janssenn B. Development of a disease specific version of the EQ-5D-5L for use in patients suffering from psoriasis: lessons learned from a feasibility study in the UK. *Value in Health. 2013.* 16; 8: 1156-62.

Szende A, Oppe M, Devlin N. EQ-5D value set. *Springer*. 2007

Szende A, Schramm W, Flood E, Larson P, Gorina E, Rentz AM, et al. Health-related quality of life assessment in adult haemophilia patients: a systematic review and evaluation of instruments. *Haemophilia.* 2003. 9; 6: 678-87.

Tabachnick BG, Fidell LS. Using multivariate statistics. *Ally and Bacon 4th ed: New York.* 2001.

Tesio L. Measuring behaviours and perceptions: Rash analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine*. 2003. 35: 105-15.

Thomson B, Daniel LG. Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement.* 1996. 56: 197-208.

Tolley K, Goad C, Yi Y, Maroudas P, Haiderali A, Thompson G. Utility elicitation study in the UK general public for late stage chronic lymphocytic leukaemia. *European Journal of Health Economics*. 2013. 14: 749-59.

Torrance GW. Multi-attribute utility theory as a method of measuring social preferences for health states in long term care. In: Kane RL, Kane RA. Values in long term care. *Lovington Books, DC Health and Co.* 1982. 127-56.

Torrance G. Measurement of health state utility for economic appraisal. *Journal of Health Economics.* 1986. 5: 1-30.

Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Economic Planning Science*. 1976. 10: 129-36.

Torrance GW. Utility approach to measuring health related quality of life. *Journal of Chronic Diseases.* 1987. 40; 6: 593-600.

Torrance GW, Feeny D. Utilities and quality adjusted life years. *International Journal of Technology Assessment in Health Care.* 1989. 5: 559-75.

Torrance GW, Feeny D, Furlong W. Visual Analogue Scales: do they have a role in the measurement of preferences for health states? *Medical Decision Making.* 2001. 21; 4: 329-34.

Torrance GW, Fenny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. A multi-attribute utility function for a comprehensive health status classification system: Health Utilities Mark II. *Medical Care.* 1996. 34: 702-22.

Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders. *Value In Health.* 2012. 15: 118.

Tussing AD, Wojtowycz MA. Physician-induced demand by Irish GPs. *Social Science and Medicine.* 1986. 23: 851-60.

TLV. General guidelines for economic evaluations from the Pharmaceutical Benefits Board. 2003. Available at: http://www.tlv.se/Upload/English/Guidelines-for-economic-evaluations-LFNAR-2003-2.pdf

Van Doorslaer, E, Wagstaff A, Rutten, F. Equity in the Finance and Delivery of Health Care: An International Perspective. *Oxford University Press: Oxford*. 1993.

Van Hout B, Janssen MF, Feng YS, Kohlmann S, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS. Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health.* 2012; 15; 5: 708-15.

Van Nooten FE, Koolman X, Brouwer WBF. The influence of subjective life expectancy on health state valuations using a 10 year TTO. *Health Economics.* 2009. 18: 549–58.

Velicer WF. An empirical comparison of the simi- larity of principal component, image, and factor patterns. *Multivariate Behavioral Research*. 1977. 12: 3-22.

Velicer WF, Jackson DN Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*. 1990. 25: 1-28.

Velicer WF, Jackson DN. Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research*. 1990. 25: 97-114.

Velicer WF, Peacock AC, Jackson, DN. A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research*. 1982. 17: 371-88.

Versteegh MM, Attema AE, Oppe M, Devlin NJ, Stolk EA. Time to tweak the TTO: results from a comparison of alternative specifications of the TTO. *European Journal of Health Economics.* 2013.

Viney R, Norman R, Brazier J, Cronin P, King MT, Ratcliffe J, Street D. An Australian discrete choice experiment to value EQ-5D health states. *Health Economics.* 2014. 23: 729-42.

Wagstaff A. QALYs and the equity-efficiency trade-off. *Journal of Health Economics.* 1991. 10: 21-41.

Ware, J. Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume 1. Model of Health and Methodology. *RAND: Santa Monica, CA.* 1980.

Ware JE, Brook RH, Davies AR, Lohr KR. Choosing measures of health status for individuals in the general population. *RAND: Santa Monica: CA.* 1981.

Wakker P, Stiggelbout A. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making.* 1995. 15: 180–6.

Walker S, Schulper M, Drummond M. The methods of cost-effectiveness analysis to inform decisions about the use of health care interventions and programs. In: Glied S and Smith PC. The Oxford Hanbook of Health Economics. *Oxford University Press: Oxford.* 2011.

Walters SJ. Quality of Life Outcomes in Clinical Trials and Healthcare Evaluation: a Practical Guide to Analysis and interpretation. *John Wiley and Sons: Chichester, UK*. 2009.

Whitehurst DG, Noonan VK, Dvorak MF, Bryan S. A review of preference-based health-related quality of life questionnaires in spinal cord injury research. *Spinal Cord.* 2012. 50; 9: 646-54.

Whynes DK. Does the correspondence between EQ-5D-3L health state description and VAS score vary by medical condition? *Health and Quality of Life Outcomes*. 2013. 11: 155.

Wilkin D, Hallam L, Doggett LA. Measures of need and outcome for primary health care. *Oxford Medical Press: Oxford.* 1992.

Wilson RN. The Sociology of Health: An introduction. *Random House: New York.* 1970

Wilson IB, Cleary PD. Linking clinical variables with health related quality of life. A conceptual model of patient outcomes. J*ournal of American Medical Association.* 1995. 273; 1: 59-65.

Wood PHN. Health and Disease and Its Importance for Models Relevant to Health Research. In Nizetic BZ, Pauli HG, Svensson PG. Scientific Approaches to Health and Health Care. *World Health Organization: Copenhagen, Denmark*. 1986. 57-70.

Wooldridge JM. Introductory econometrics: a modern approach. *Mason 4th ed*. *South-Western Cengage Learning: OH*. 2009.

Wolfs CAG, Dirksen CD, Kessels A, Willems DCM, Verhey FRJ, Severens JL. Performance of the EQ-5D and the EQ-5D + cognition in elderly patients with cognitive impairments. *Health and Quality of Life Outcomes*. 2007. 5; 33.

Wolfson AD, Sinclair AJ, Bombardier C, McGreer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA, editors. Values and long term care. *Lexington, Mass. Lexington Books.* 1982. 191–214.

World Health Organization. Constitution of the World Health Organization. *World Health Organization: Geneva.* 1948.

World Health Organization. International classification of Impairments, Disability and Handicaps. *World Health Organization: Geneva.* 1980.

World Health Organization. International Classification of Diseases. *World Health Organization: Geneva.* 2010. Available at: http://www.who.int/classifications/icd/en/

Wu AW, Hanson KA, Harding G, Haider S, Tawadrous M, Khachatryan A, et al. Responsiveness of the MOS-HIV and EQ-5D in HIV-infected adults receiving antiretroviral therapies. *Health and Quality of Life Outcomes.* 2013. 12: 11.

Yang Y, Brazier JE, Tsuchiya A. Effect of adding a sleep dimension to the EQ5D descriptive system: A "bolt-on" experiment. *Medical Decision Making*. 2014 (b). 34: 42-53.

Yang Y, Longworth L, Brazier J. An assessment of validity and responsiveness of generic measures of health-related quality of life in hearing impairment. *Quality of Life Research.* 2013.

Yang Y, Brazier J, Longworth L. EQ5D in skin conditions: an assessment of validity and responsiveness. *European Journal of Health Economics*. 2014 (a).

Yang Y, Rowen D, Brazier J, Tsuchiya A, Young T, Longworth L. An Exploratory Study to Test the Impact on Three "Bolt-On" Items to the EQ-5D. *Value in Health.* 2015. 18: 52-60.

Young T, Yang Y, Brazier J, Tsuchiya A. Using Rash analysis to aid the construction of preference based measures from existing quality of life instruments. *Paper presented at the Health Economics Study Group meeting.* Newcastle, *UK*. 2005.

Appendices Chapter 3

Table I – Searching strategy for Medline

Search	*Searching terms*
#1	validity OR responsiveness OR "psychometric characteristic$" OR "psychometric aspect$" OR "psychometric propert$"
#2	"Preference based instrument$" OR "preference based measure$" OR "multi-attribute utility instrument$" OR "generic adj instrument$" OR "multi adj instrument$" OR "patient reported outcome$" OR "PROMS" OR "PROM"
#3	euroqol OR "euro qol" OR "eq5d" OR "eq 5d" OR "eq-5d" OR "euro adj qol" OR "eur adj qual" OR "eq adj 5d[tw]"
#4	sf6D OR "sf 6D" OR "short form 6D" OR "shortform 6D" OR "sf six D" OR "sfsixD" OR "shortform six D" OR "short form sixD" OR "sf-6d" OR 6d OR 6-d OR "6 dimension[tw]"
#5	"hui3" OR "hui 3" OR "health utilities index mark 3" OR "health utilities mark three" OR "hui III" OR "hui mk III" OR huiIII[tw]
#6	"15D instrument" OR "fifteen D instrument" OR "fifteen dimension instrument" OR "15 dimension instrument" OR "15dimension instrument"
#7	"Assessment of Quality of Life" OR AQOL OR "Assessment-of-Quality-of-Life"
#8	"Quality of life"
#9	systematic[sb] OR meta-analysis[pt] OR meta-analysis as topic[mh] OR meta-analysis[mh] OR meta analy*[tw] OR metanaly*[tw] OR metaanaly*[tw] OR met analy*[tw] OR integrative research[tiab] OR integrative review*[tiab] OR integrative overview*[tiab] OR research integration*[tiab] OR research overview*[tiab] OR collaborative review*[tiab] OR collaborative overview*[tiab] OR systematic review*[tiab] OR technology assessment*[tiab] OR technology overview*[tiab] OR "Technology Assessment, Biomedical"[mh] OR HTA[tiab] OR HTAs[tiab] OR comparative efficacy[tiab] OR comparative effectiveness[tiab] OR outcomes research[tiab] OR indirect comparison*[tiab] OR ((indirect treatment[tiab] OR mixed-treatment[tiab]) AND comparison*[tiab]) OR Embase*[tiab] OR

	Cinahl*[tiab] OR systematic overview*[tiab] OR methodological overview*[tiab] OR methodologic overview*[tiab] OR methodological review*[tiab] OR methodologic review*[tiab] OR quantitative review*[tiab] OR quantitative overview*[tiab] OR quantitative synthes*[tiab] OR pooled analy*[tiab] OR Cochrane[tiab] OR Medline[tiab] OR Pubmed[tiab] OR Medlars[tiab] OR handsearch*[tiab] OR hand search*[tiab] OR meta-regression*[tiab] OR metaregression*[tiab] OR data synthes*[tiab] OR data extraction[tiab] OR data abstraction*[tiab] OR mantel haenszel[tiab] OR peto[tiab] OR der-simonian[tiab] OR dersimonian[tiab] OR fixed effect*[tiab] OR "Cochrane Database Syst Rev"[Journal:__jrid21711] OR "health technology assessment winchester, england"[Journal] OR "Evid Rep Technol Assess (Full Rep)"[Journal] OR "Evid Rep Technol Assess (Summ)"[Journal] OR "Int J Technol Assess Health Care"[Journal] OR "GMS Health Technol Assess"[Journal] OR "Health Technol Assess (Rockv)"[Journal] OR "Health Technol Assess Rep"[Journal]
#10	#2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8
#11	#1 AND #10
#12	#9 AND #11

Table II -AMSTAR Modified version

Question	Score
Was an apriori design provided?	0.5 points
Was there duplicate study selection and data extraction?	1 point
Was a comprehensive literature search performed?	2 points
Was a list of included studies provided?	0.5 points
Where the characteristics of the included studies provided?	1.5 points
Was the scientific quality of the included studies assessed and documented?	2 points
Was the scientific quality of the included studies used appropriately in formulating conclusions?	2 points
Was the conflict of interest included?	0.5 points

Minimum score 0

Maximum score 10

Table III - Scores for original and modified AMSTAR checklist

Disease area	*Report*	*AMSTAR modified score*	*AMSTAR original score*
Autoimmune system	Castelino	2,5	4/11
	Holloway	3,5	4/11
Cardiovascular system	Dyer	4,5	5/11
Ear	Yang	7	7/11
Endocrine, nutritional and metabolic diseases	Janssen	5,5	7/11
	Speight	0,5	2/11
Eye	Tosh	7	7/11
Genitourinary system	Davis and Wailoo	5,5	6/11
	Wu	5,5	6/11
Gynaecological problems	Sanghera	2,5	3/11
Haematological problems	Szende	4,5	6/11
Musculoskeletal system	Bansback	4,5	5/11
	DeVine	3	4/11
	Hill	4,5	5/11
	Whitehurst	5,5	6/11
Mental health	Brazier	5	5/11
	Papaioannou	5	6/11
	Papaioannou	7	7/11
	Peasgood	6,5	6/11
	Hounsome	4	5/11
Neoplasm	Longworth	7	7/11

	Pickard	4	5/11
Nervous system	Kuspinar and Mayo	8,5	8/11
Nose	Linder	9,5	8/11
Others	Ching	0	1/11
	Derrett	3	4/11
	Haywood	3	3/11
Respiratory system	Petrillo	2,5	4/11
	Pickard	5,5	8/11
Skin and subcutaneous tissues	Yang	7	7/11

Note: Cutoffs used for quality assessment are excellent quality ≥7.5, good quality ≥5, poor quality <5.

Table IV– List of included studies

Studies
Bansback N, Ara R, Karnon J, Anis A. Economic evaluations in rheumatoid arthritis: a critical review of measures used to define health States. Pharmacoeconomics. 2008. 26; 5:395-408.
Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess. 2014. 18; 34: 1-188.
Castelino, M, Abbott J, McElhone K, The LS. Comparison of the psychometric properties of health-related quality of life measures used in adults with systemic lupus erythematosus: a review of the literature. Rheumatology. 2013. 52; 684.
Ching S, Thoma A, McCabe RE, Antony MM. Measuring outcomes in aesthetic surgery: a comprehensive review of the literature. Plast Reconstr Surg. 2003. 111; 1:469-80.
Davis S, Wailoo A. A review of the psychometric performance of the EQ-5D in people with urinary incontinence. Health Qual Life Outcomes. 2013. 18; 11.
Derrett S, Black J, Herbison GP. Outcome after injury-a systematic literature search of studies using the EQ-5D. J Trauma. 2009. 67; 4: 883-890.
DeVine J, Norvell DC, Ecker E, Fourney DR, Vaccaro A, Wang J, et al. Evaluating the correlation and responsiveness of patient-reported pain with function and quality-of-life outcomes after spine surgery. Spine (Phila Pa 1976). 2011. 36; 21: S69-74.
Dyer MTD, Goldsmith KA, Sharples LS, Buxton MJ. A review of health utilities using the EQ5D in studies of cardiovascular disease. Health and Quality of Life Outcomes. 2010.
Haywood KL, Garratt AM, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. Qual Life Res. 2005. 14; 7:1651-68.
Hill MR, Noonan VK, Sakakibara BM, Miller WC, SCIRE Research Team. Quality of life instruments and definitions in individuals with spinal cord

injury: a systematic review. Spinal Cord. 2010. 48; 6:438-450.

Holloway L, Humphrey L, Heron L, Pilling C, Kitchen H, Hojbjerre L, et al. Patient-reported outcome measures for systemic lupus erythematosus clinical trials: a review of content validity, face validity and psychometric performance. Health Qual Life Outcomes. 2014. 22; 12.

Hounsome N, Orrell M, Edwards RT. EQ-5D as a quality of life measure in people with dementia and their carers: evidence and key issues. Value Health. 2011. 14; 2: 390-99.

Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with Type 2 diabetes mellitus. Diabet Med. 2011. 28; 4: 395-413.

Kuspinar A, Mayo NE. A review of the psychometric properties of generic utility measures in multiple sclerosis. Pharmacoeconomics. 2014. 32; 8: 759-73.

Linder JA, Singer DE, Ancker M, Atlas SJ. Measures of health-related quality of life for adults with acute sinusitis. A systematic review. J Gen Intern Med. 2003. 18; 5: 390-401.

Longworth L, Yang Y, Young T, Mulhern B, Hernandez Alava M, Mukuria C, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. Health Technol Assess. 2014. 18; 9:1-224.

Papaioannou D, Brazier J, Parry G. How to measure quality of life for cost effectiveness analysis in personality disorder? A systematic review. HEDS discussion paper 13/02. 2013.

Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value Health. 2011. 14; 6: 907-920.

Peasgood T, Brazier J, Papaioannou D. A systematic review of the validity and responsiveness of EQ5D and SF6D for depression and anxiety. HEDS discussion paper 12/15 (unpublished). 2012

Petrillo J, van Nooten F, Jones P, Rutten-van Molken M. Utility estimation in chronic obstructive pulmonary disease: a preference for change? Pharmacoeconomics. 2011. 29; 11:917-32.

Pickard AS, Wilke CT, Lin HW, Lloyd A. Health utilities using the EQ-5D in studies of cancer. Pharmacoeconomics. 2007. 25; 5: 365-84.

Pickard AS, Wilke C, Jung E, Patel S, Stavem K, Lee TA. Use of a preference-based measure of health (EQ-5D) in COPD and asthma. Respir Med. 2008. 102; 4: 519-36.

Sanghera S, Frew E, Kai J, Gupta J, Elizabeth Roberts T. An assessment of economic measures used in menorrhagia: a systematic review. Soc Sci Med. 2013. 98: 149-53.

Speight J, Reaney MD, Barnard KD. Not all roads lead to Rome-a review of quality of life measurement in adults with diabetes. Diabet Med. 2009. 26; 4: 315-27.

Szende A, Schramm W, Flood E, Larson P, Gorina E, Rentz AM, et al. Health-related quality of life assessment in adult haemophilia patients: a systematic review and evaluation of instruments. Haemophilia. 2003. 9; 6: 678-87.

Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders. Value In Health.2012.15: 118.

Whitehurst DG, Noonan VK, Dvorak MF, Bryan S. A review of preference-based health-related quality of life questionnaires in spinal cord injury research. Spinal Cord. 2012. 50; 9: 646-54.

Wu AW, Hanson KA, Harding G, Haider S, Tawadrous M, Khachatryan A, et al. Responsiveness of the MOS-HIV and EQ-5D in HIV-infected adults receiving antiretroviral therapies. Health Qual Life Outcomes. 2013. 12; 11.

Yang Y, Longworth L, Brazier J. An assessment of validity and responsiveness of generic measures of health-related quality of life in hearing impairment. Quality of Life Research. 2013.

Yang Y, Brazier J, Longworth L. EQ5D in skin conditions: an assessment of validity and responsiveness. European Journal of Health Economics. 2014.

Table V – List of excluded studies

Studies
Ades AE, Lu G, Madan JJ. Which health-related quality-of-life outcome when planning randomized trials: Disease-specific or generic, or both? A common factor model. Value in Health. 2013. 16; 1: 185-94.
Adobor RD, Rimeslatten S, Keller A, Brox JI. Repeatability, reliability, and concurrent validity of the scoliosis research society-22 questionnaire and EuroQol in patients with adolescent idiopathic scoliosis. Spine. 2010. 35: 206.
Al Sayah F, Ishaque S, Lau D, Johnson JA. Health related quality of life measures in Arabic speaking populations: a systematic review on cross-cultural adaptation and measurement properties. Qual Life Res. 2013. 22; 1: 213-29.
Alsanafi S, Werth VP. Evaluating quality of life in dermatomyositis. Expert Review of Dermatology. 2010. 5; 6: 605-9.
Angst F, Stucki G, Aeschlimann A. Quality of life assessment in osteoarthritis. Expert Review of Pharmacoeconomics and Outcomes Research. 2003. 3; 5: 623-36.
Arrington R, Cofrancesco J, Wu AW. Questionnaires to measure sexual quality of life. Qual Life Res. 2004. 13; 10: 1643-58.
Ashcroft DM, Li Wan Po A, Williams HC, Griffiths CE. Quality of life measures in psoriasis: a critical appraisal of their quality. J Clin Pharm Ther. 1998. 23; 5: 391-8.
Au Eong KG, Chan EW, Luo N, Wong SH, Tan NW, Lim TH, Wagle AM. Validity of EuroQOL-5D, time trade-off, and standard gamble for age-related macular degeneration in the Singapore population. Eye. 2012. 26: 379.
Augestad LA, Rand-Hendriksen K, Kristiansen IS, Stavem K. Learning effects in time trade-off based valuation of EQ-5D health states. Value in Health. 2012. 15; 2: 340-45.
Augustin M, Langenbruch AK, Gutknecht M, Radtke MA, Blome C. Quality of Life Measures for Dermatology: Definition, Evaluation, and Interpretation. Current Dermatology Reports. 2012. 1; 3: 148-59.
Augustin M, Langenbruch AK, Herberger K, Baade K, Goepel L, Blome C. Quality of life measurement in chronic wounds and inflammatory skin diseases: Definitions, standards and instruments. Wound Medicine. 2014. 5: 29-38.
Balioussis C, Hitzig SL, Flett H, Noreau L, Craven BC. Identifying and classifying quality of life tools for assessing spasticity after spinal cord injury. Topics in Spinal

Cord Injury Rehabilitation. 2014. 20; 3: 208-24.

Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Comparing the performance of the EQ-5D and SF-6D when measuring the benefits of alleviating knee pain. Cost Effectiveness and Resource Allocation. 2009. 7: 12.

Beresniak A, Russell AS, Haraoui B, Bessette L, Bombardier C, Duru G. Advantages and limitations of utility assessment methods in rheumatoid arthritis. J Rheumatol. 2007. 34; 11: 2193-2200.

Blome C, Beikert FC, Rustenbach SJ, Augustin M. Mapping DLQI on EQ-5D in psoriasis: transformation of skin-specific health-related quality of life into utilities. Arch Dermatol Res. 2013. 305; 197.

Boling W, Fouladi RT, Basen-Engquist K. Health-related quality of life in gynecological oncology: instruments and psychometric properties. Int J Gynecol Cancer. 2003. 13; 1: 5-14.

Brazier JE, Connell J, O'Cathain A. DO EQ-5D and SF-6D ask the right questions in mental health? A content validation using interviews with patients. Value in Health. 2014. 17; 3: A194.

Brazier JE, Green C, Kanis JA. A systematic review of health state utility values for osteoporosis-related conditions. Osteoporosis Int. 2002. 13; 10: 768-76.

Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, et al. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess. 2012. 16; 32:1-114.

Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2010. 11; 2: 215-25.

Brennan DS. Oral Health Impact Profile, EuroQol, and Assessment of Quality of Life instruments as quality of life and health-utility measures of oral health. Eur J Oral Sci. 2013. 121; 188.

Buitinga L, Braakman-Jansen LM, Taal E, Kievit W, Visser H, van Riel PL, van de Laar MA. Comparative responsiveness of the EuroQol-5D and Short Form 6D to improvement in patients with rheumatoid arthritis treated with tumor necrosis factor blockers: results of the Dutch Rheumatoid Arthritis Monitoring registry. Arthritis Rheum. 2013. 64: 826.

Chopra I, Kamal KM. A systematic review of quality of life instruments in long-term breast cancer survivors. Health Qual Life Outcomes. 2012. 31; 10.

Chow MY, Morrow AM, Cooper Robbins SC, Leask J. Condition-specific quality of life questionnaires for caregivers of children with pediatric conditions: a systematic review. Qual Life Res. 2013. 22; 8: 2183-2200.

Cleland J, Gillani R, Bienen EJ, Sadosky A. Assessing dimensionality and responsiveness of outcomes measures for patients with low back pain. Pain Pract. 2011. 11; 1: 57-69.

Connell J, O'Cathain A, Brazier J. Measuring quality of life in mental health: Are we asking the right questions?. Soc Sci Med. 2014. 20.

Danchenko N, Rive B, Pendlebury S, Abetz L. Assessing health-related quality of life (HRQOL) in major depressive disorder (MDD): Do commonly used instruments meet EMA requirements for psychometric properties and content relevance. Value in Health. 2013. 16; 3:A64.

Danquah FV, Wasserman J, Meininger J, Bergstrom N. Quality of life measures for patients on hemodialysis: a review of psychometric properties. Nephrol Nurs J. 2010. 37; 3: 255-69.

Doctor JN, Bleichrodt H, Lin HJ. Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Making. 2010. 30; 1: 58-67.

Gerhards SAH, Huibers MJH, Theunissen KATM, De Graaf LE, Widdershoven GAM, Evers SMAA. The responsiveness of quality of life utilities to change in depression: A comparison of instruments (SF-6D, EQ-5D, and DFD). Value in Health. 2011. 14; 5: 732-39.

Golomb BA, Vickrey BG, Hays RD. A review of health-related quality-of-life measures in stroke. Pharmacoeconomics. 2001. 19; 2: 155-85.

Heintz E, Wirehn AB, Peebo BB, Rosenqvist U, Levin LA. QALY weights for diabetic retinopathy--a comparison of health state valuations with HUI-3, EQ-5D, EQ-VAS, and TTO. Value In Health. 2012. 15: 475.

Hutter BO, Wurtemberger G. Functional capacity (dyspnea) and quality of life in patients with chronic obstructive lung disease (COPD): instruments of assessment and methodological aspects. Pneumologie. 1999. 53; 3: 133-42.

Janssens L, Gorter JW, Ketelaar M, Kramer WL, Holtslag HR. Health-related quality-of-life measures for long-term follow-up in children after major trauma. Qual Life Res. 2008. 17; 5: 701-13.

Johansson P, Agnebrink M, Dahlstrom U, Brostrom A. Measurement of health-related quality of life in chronic heart failure, from a nursing perspective-a review of the literature. Eur J Cardiovasc Nurs. 2004. 3; 1: 7-20.

Johnsen LG, Hellum C, Nygaard OP, Storheim K, BroxJI, Rossvoll I, Leivseth G, Grotle M. Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease. BMC Musculoskeletal Disorders. 2013. 14: 148.

Jones GL, Hall JM, Balen AH, Ledger WL. Health-related quality of life measurement in women with polycystic ovary syndrome: a systematic review. Hum Reprod Update. 2008. 14; 1: 15-25.

Koivunen K, Sintonen H, Lukkarinen H. Properties of the 15D and the Nottingham Health Profile questionnaires in patients with lower limb atherosclerotic disease. Int J Technol Assess Health Care. 2007. 23; 3: 385-91.

Kontodimopoulos N, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value in Health. 2009. 12; 8: 1151-57.

Lee EH, Kim CJ, Cho SY, Chae HJ, Lee S, Kim EJ. Monitoring the use of health-related quality of life measurements in Korean studies of patients with diabetes. J Korean Acad Nurs. 2011. 41; 4: 558-67.

Lee WJ, Song KH, Noh JH, Choy YJ, Jo MW. Health-related quality of life using the EuroQol 5D questionnaire in Korean patients with type 2 diabetes. J Korean Med Sci. 2012. 27; 255.

Luscombe FA. Health-related quality of life measurement in type 2 diabetes. Value in Health. 2000. 3; 1:15-28.

Millier A, Clay E, Chauhan D, Toumi M. Health-related quality of life patient-reported outcomes in schizophrenia. Value in Health. 2012. 15; 7: A342.

Ortiz Z, Shea B, Garcia Dieguez M, Boers M, Tugwell P, Boonen A, et al. The responsiveness of generic quality of life instruments in rheumatic diseases. A systematic review of randomized controlled trials. J Rheumatol. 1999. 26; 1: 210-16.

Papaioannou D, Peasgood T, Brazier J, Parry G. The validity of the EQ-5D, SF-6D, SF-36 and SF-12 in mental health conditions: A systematic review. Value in Health. 2011. 14; 7: A239.

Patel KK, Veenstra DL, Patrick DL. A review of selected patient-generated outcome measures and their application in clinical trials. Value in Health. 2003. 6; 5: 595-603.

Reaney MD, Martin C, Speight J. Understanding and Assessing the Impact of

Alcoholism on Quality of Life: A Systematic Review of the Content Validity of Instruments Used to Assess Health-Related Quality of Life in Alcoholism. Patient. 2008. 1; 3: 151-63.

Rowen D, Young T, Brazier J, Gaugris S. Comparison of generic, condition-specific, and mapped health state utility values for multiple myeloma cancer. Value In Health. 2012. 15: 1059.

Schmier JK, Halpern MT, Higashi MK, Bakst A. The quality of life impact of acute exacerbations of chronic bronchitis (AECB): a literature review. Qual Life Res. 2005.14; 2: 329-47.

Scholzel-Dorenbos CJ, van der Steen MJ, Engels LK, Olde Rikkert MG. Assessment of quality of life as outcome in dementia and MCI intervention trials: a systematic review. Alzheimer Dis Assoc Disord. 2007. 21; 2: 172-78.

Shaw JW, Joish VN, Coons SJ. Onychomycosis: health-related quality of life considerations. Pharmacoeconomics. 2002. 20; 1: 23-36.

Smith AB, Cocks K, Taylor M, Parry D. Responsiveness of the EQ-5D in oncology: A meta-analysis. Value in Health. 2013. 16; 7: A323-A324.

Soer R, Reneman MF, Speijer BL, Coppes MH, Vroomen PC. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. Spine Journal. 2012. 12: 1035.

Street J, Berven S, Fisher C, Ryken T. Health related quality of life assessment in metastatic disease of the spine: a systematic review. Spine (Phila Pa 1976). 2009. 15; 34: S128-34.

Takemoto MLS, Fernandes RA, Cukier FN, Cruz RB, Takemoto MMS, Santos PML, et al. The measurement and valuation of health status using EQ-5D in Brazil: A systematic review. Value in Health. 2011. 14; 7: A404.

Teckle P, Peacock S, McTaggart-Cowan H, van der Hoek K et al. The ability of cancer-specific and generic preference-based instruments to discriminate across clinical and self-reported measures of cancer severities. Health and Quality of Life Outcomes. 2011. 9: 106.

Tordrup D, Mossman J, Kanavos P. Responsiveness of the EQ-5D to clinical change: is the patient experience adequately represented? Int J Technol Assess Health Care. 2014. 30; 1: 10-19.

Tordrup D, Mossman J, Kanavos P. Shortcomings of EQ-5D in a value-based pricing framework. Value in Health. 2013. 16; 3: A10-A11.

Torrecillas IV, Sarabia FN. Arthrosis and quality of life. DOLOR. 1999 14; 2: 121-130.

Turner N, Campbell J, Peters TJ, Wiles N, Hollinghurst S. A comparison of four different approaches to measuring health utility in depressed patients. Health and Quality of Life Outcomes. 2013. 11; 81.

Van Hanswijck de Jonge P, Lloyd A, Horsfall L, Tan R, O'Dwyer PJ. The measurement of chronic pain and health-related quality of life following inguinal hernia repair: a review of the literature. Hernia. 2008. 12; 6: 561-69.

Whitehurst DG, Engel L, Bryan S. Short Form health surveys and related variants in spinal cord injury research: a systematic review. J Spinal Cord Med. 2014. 37; 2: 128-38.

Wilke CT, Pickard AS. Test-retest reliability of the EQ-5D visual analog scale across populations and conditions. Value in Health. 2009. 12; 3: A30.

Table VI – Known group validity main results

Report	Condition	Quality	GPBM	N studies	Groups	Main results
Brazier	Bipolar disorder	Good	EQ5D	2	Severity	Two studies support validity (p N/R)
				1	Patients /general population	One study does not support validity (p N/R)
Papaioannou (a)	Personality disorder	Good	EQ5D	1	Different types of disease/disorder	In one study the number of personality disorders (e.g. narcissistic; obsessive-compulsive) have a significant effect on EQ5D scores (p=0.000). When controlling for the number of disorders in linear regression, only depressive personality disorder maintained a unique statistical effect on QoL (p=0.03)
				1	Patients /general population	In one study individuals with complex personality disorders had much lower EQ5D scores in comparison to the general population (0.54 vs. 0.85) (p N/R)
Papaioannou (b)	Schizophrenia	Good	EQ5D	1	Severity	In one study EQ5D scores differed according to the severity of the disease, with clinically significant results (defined as >0.03) in line with those of the other instrument (p N/R).

Peasgood	Depression/ Anxiety	Good	EQ5D	3	Severity	In one study EQ5D presented significant differences in values according to the severity levels measured by the severity of illness scale (p N/R). Similar results were found in another study, but differences were not statistically significant between moderate and severe cases. In a third study EQ5D scores decreased as anxiety symptom severity increased (measured by Hospital Anxiety and Depression Scale), with statistically significant results for all levels but normal versus mild.
				3	Different types of disease/disorder	In one study EQ5D distinguished between single and recurrent depression but differences were not statistically significant while results were statistically significant for the physical component of SF36. In one study EQ5D values were different depending on the type of depressive disorder (p N/R). In another EQ5D values differed according to whether patients had major depressive disorder, anxiety, both or none (p N/R).

			EQ5D, SF6D	2	Severity	In one study both measures showed expected patterns comparing groups. EQ5D presented larger drops in utility in the most severe quartile, which were not found in SF6D. Standard deviations were greater in EQ5D than in the SF6D. Difference between mean and median greater for EQ5D. In the other study EQ5D and SF6D were sensitive to different severity levels, although SF6D showed lower index values for moderately severe compared to severe patients (p N/R). One study analysed the relative efficiency statistic of the two instruments for different levels of self-rated health, finding SF6D to be more efficient by 29% to 423% (p not statistically significant).
			EQ5D, HUI 3	1	Severity	Both EQ5D and HUI 3 presented changes in the expected direction at increasing levels of anxiety and compared favourably to other clinical measures, although HUI3 appeared more sensitive at increasing levels of severity. All differences were statistically significant.
			SF6D	2	Different types of disease/diso	In one study, EQ5D differentiated between anxiety disorder, depressive disorder, anxiety and depressive disorder and no disorders

					rder	patients (p N/R). In another study similar results were found for patients with different types of disorders (e.g. panic, depressive, anxiety, social phobia etc.) (p N/R).
Davis and Wailoo	Urinary Incontinence	Good	EQ5D	3	Severity	In two studies EQ5D direction of change was consistent across groups and with clinical expectations, in one with p>0.05 and in the other with p N/A. In a third study, direction of change was consistent across groups and consistent with clinical expectation for the two least severe groups, but p was not statistically significant.
				1	Different types of disease/diso rder	In one study direction of change was consistent across instruments for patients with general incontinence, stress incontinence or no incontinence (p=N/R).
			EQ5D, SF6D, AQoL	1	Severity	In one study instruments presented direction of change consistent across groups and with clinical expectation at p<0.0001.
Derrett	Injuries	Poor	EQ5D	3	Severity	In two studies both the EQ-5D and the Nottingham Health Profile discriminated between patients with displaced and un-displaced fractures and EQ-5D was considered

						almost as effective alone as in conjunction with the larger Nottingham Profile (p N/R). In the other study EQ-5D could correctly classify un-displaced fractures from more debilitating displaced femoral neck fractures with 74.5% accuracy, which was greater than Nottingham Health Profile Accuracy (p N/R).
Janssen	Diabetes type 2	Good	EQ5D	1	Severity	In one study EQ5D index was 0.042 (95% CI 0.003-0.088) higher for each 0.1 increment in Kt/V (hemodialysis and peritoneal dialysis treatment adequacy), which was both statistically and clinically significant.
				1	Patients /general population	In one study the EQ5D discriminated well between patients according to presence or absence of diabetes (p N/R).
				1	Type of disease/disorder	In one study the instrument was able to discriminate between different histories of hypoglycaemia (p N/R).
			EQ5D, SF6D	1	Patients /general population	In one study the EQ5D index and VAS were able to discriminate between patients with diabetes and patients without diabetes. The SF6D was not able to discriminate between

						these groups (p N/R)
Kuspinar and Mayo	Multiple Sclerosis	Excellent	EQ5D	3	Severity	In one study the EQ5D was able to distinguish just for some of the levels of severity assessed through the expanded disability status scale (utility score for level 3 higher than for level 4) (p N/R). In other two studies it was found that EQ5D lacked discriminative ability for those group of patients that were wheelchair bounded (p N/R).
			AQoL	2	Severity	In one study the AQoL was able to differentiate patients with different levels of pain intensity (p N/R). In another between mildly, moderately and severely disabled patients (p N/R).
			SF6D, HUI3	1	Severity	In one study the HUI 3 demonstrated known group validity by being able to differentiate between mild, moderately and severely disabled multiple sclerosis patients. Although the SF6D was able to differentiate between mildly and moderately disabled patients, it was unable to differentiate between the more severe patient groups. A flattening of utility scores beyond moderate disability was observed (p N/R)

Tosh	Visual impairment	Good	EQ5D	10	Severity	In two studies EQ5D decreased appropriately according to severity groups although decrements were not statistically significant in one and not reported in the other group. In a third study EQ5D decreased with increasing glaucoma damage, but the difference between groups were not statistically significant except for the most severe group. In a fourth study on age related macular degeneration EQ5D showed inconsistencies, with normal visual acuity patients having a worst mean utility than mild, moderate, severe and near blind patients. This inconsistency was not seen in the visual functional questionnaire 20/25. For patients with diabetic retinopathy, two studies found statistically significant difference between the two extreme groups but differences between neighbouring groups were not significant and frequently inconsistent. In another study EQ5D showed appropriate but non-significant difference between low and high visual field groups, but an inconsistent and non-significant difference in the EQ5D between low and high visual acuity severity groups. In one study on patients with cataracts showed appropriate but

						non-significant changes in EQ5D between the first and second eye surgery groups. Other two studies could not find significant evidence to support the association between the degree of Visual acuity and EQ5D in patients with cataracts.
				12	Patients/ general population	In three studies EQ5D showed an appropriate and statistically significant reduction in HRQoL for people with age related macular degeneration compared to the general population while in one appropriate but not statistically significant. Other three studies found statistically significant difference between patients with and without conjunctivitis. Among them, in one study it is mentioned that the pain dimension appears to be the only dimension to show a statistical difference, while in another that all dimension were significant except mobility. In other two studies patients with endophtalmitis and patients with cytomegalovirus were found to have appropriate but non-statistically

						significant differences compared to the general population. In another study patients with cataract and general population registered significant differences across all dimensions except pain. An additional two studies found statistically significant and appropriate differences between groups of patients with unspecified blindness/ visual impairment and general population.
			EQ5D, SF6D, HUI3	1	Severity levels	One study found a consistent relationship between different severity assessed through visual acuity or contrast sensitivity and utility values for SF6D and HUI III, but not for the EQ5D
Yang (a)	Hearing impairment	Good	HUI 3	4	Severity levels	In three studies HUI 3 differentiated between levels of severity but p values were not reported. In one did not differentiate between unilateral and bilateral implantation (p value not reported) and this was consistent with VAS scores.
			EQ5D, HUI 3	1	Severity levels	EQ5D failed to detect significant differences by hearing loss severity groups where HUI 3 detected them (p N/R)

Yang (b)	Skin condition	Good	EQ5D	2	Severity levels	In two studies statistically significant differences (in one p<0.01, in the other p<0.05) according to severity groups, similar to other measures.
				7	Patients/ general population	In six studies difference were statistically significant in patients with psoriasis (three), nail psoriasis (one), hand eczema (one) and acne (one) and the general population. In one study on psoriasis results were not statistically significant. In the study on acne most EQ5D dimensions were sensitive and statistically significant, especially for anxiety depression and pain/discomfort.
Longworth	Cancer	Good	EQ5D	14	Severity levels	In one study mixed patterns were found in the EQ5D, which did not find group differences that were reported using other generic (SF12) and condition specific measures for the mildly dyskaryotic group. Mixed results were found also for breast cancer (one study) (p N/R). In two studies EQ5D could not distinguish between colon cancer groups where other instruments were able to (not clear whether statistically significant) and in one between gastric cancer groups (p N/R). In three studies

						on colon cancer the instrument was found to distinguish correctly and to report statistically significant results. In other seven (e.g. general cancer/lung cancer/lymphoma) EQ5D performed well but p values were statistically significant only in two.
				3	Different types of disease/disorder	In three studies, EQ5D could discriminate between the level of HRQOL associated with different types of cancer (p N/R in two, statistically significant in one).
				7	Patients /general population	In three studies, EQ5D differentiated between people with different types of cancer (e.g. pancreatic)/ general population (p statistically significant in one, N/A in two and N/R in one) and in two statistically significantly between people with metastases/people without metastases. In one study EQ5D usual activity dimension reported greater problems for cancer patients than for other population, but no difference was found in the other dimensions.
			HUI 3	2	Severity levels	In two studies HUI 3 was able to distinguish different severity groups, with results statistically significant in one and N/R in the

						other.
				1	Different types of disease/disorder	The HUI3 scores for the vision dimension was higher in the Hodgkin's group compared with acute lymphoblastic leukaemia (p<0.01). The difference between the emotion (p<0.01) and HRQoL (p<0.05) scores were significantly different with the Canadian group displaying higher mean scores. As expected, the differences in mean single attribute scores between acute lymphoblastic leukaemia and Hodgkin's disease patients were not statistically significant.
				1	Patients /general population	In one study HUI 3 was able to discriminate between patients and general population, with results that were statistically significant.
			EQ5D, SF6D	1	Severity levels	In one study the EQ5D and SF6D discriminated between symptom based severity groups as indicated by the number of symptoms. A higher number of symptoms resulted in lower utility scores. P values were mixed.
				1	Patients /general	In one study the EQVAS found a significant difference between groups that were not

					population	detected by the EQ5D and the SF6D (P values not statistically significant).
Petrillo	Asthma/COPD	Poor	EQ5D	1	Severity	In one study results were reported to be in line with expectations (p N/R).
				1	Patients/general population	In one study asthma was a marginally significant predictor of the model.
Pickard (a)	Asthma/COPD	Good	EQ5D	4	Severity	In one study EQ5D index (US, UK weights) was able to detect differences between GOLD stages (p<0.001). EQ5D index (UK, US algorithms) differentiated between GOLD II and III (ES=0.22 for US, 0.18 for UK) better than gold 3 and 4 (ES=0.47 for both). Two studies found similar results but p values were N/R. In one inconsistencies were found (p values N/R).
			EQ5D	3	Patients /general population	In one study asthma was a significant predictor of EQ5D index after controlling for sex, age, race/ethnicity and condition (p<0.0001). Similarly, in another study EQ5D showed that patients with COPD had lower health (no p values reported) and in a third study that asthma and COPD groups were more likely to

						report problems than general public (p<0.05).
Pickard (b)	Cancer	Good	EQ5D	1	Severity	In one study, EQ5D was able to distinguish between severity defined by age adjusted international prognostic index (p N/R)
				3	Patients /general population	All three studies reported that EQ5D distinguished between cancer and non cancer population, but p values were N/R.
Szende	Haemophilia	Good	EQ5D	1	Severity	EQ5D reflected well severity levels in hemophilia patients (p N/R).
				1	Patients /general population	HIV status was associated with substantially lower mean utility values (p N/R).
			HUI 3	1	Severity	HUI 3 detected that patients with HIV or hepatitis had lower utility values than patients without these comorbidities within the same haemophilia severity group. The HUI 3 could detect that this additional burden was mainly due to more problems in ambulation and pain. A difference between patients with or without hepatitis B or C was detected in patients with moderate and severe haemophilia. The lack of

						similar difference in mild haemophilia may reflect less hepatic injury because of smaller viral load from lower exposure to blood products
Dyer	Heart disease	Good	EQ5D	3	Severity	The EQ5D index showed excellent discrimination between patients with CCS scores from I-IV (p<0.01). Consistent results were found in other two studies, but p values were not reported.
				4	Different types of disease/disorder	All 4 studies differentiated appropriately between health condition, but did not reported p values.
				2	Patients/general population	Both studies found inconsistent results (p N/R)

Table VII – Convergent validity main results

Report	Condition	Quality	GPBM	Type of comparator	N studies	Main results
Bansback	Rheumatoid Arthritis	Poor	EQ5D, SF6D, HUI3	Functional status	1	Health assessment questionnaire is very strongly correlated with the EQ5D (r=0.61; p value N/R), the SF6D (r=0.73; p value N/R) and the HUI III (r=0.76; p value N/R).
Papaioannou (a)	Personality Disorder	Good	EQ5D	Symptoms and severity	2	In one study, EQ5D showed moderate correlations with BPDSI-IV (0.487) that were slight higher than correlations between EQVAS (0.404) and the severity measure (both statistically significant at p<0.01). In another study, EQ5D scores were moderately correlated with the Global Severity Index (0.49) (p N/R).
Papaioannou (b)	Schizophrenia	Good	EQ5D	HRQoL	4	In one study EQ5D was moderately to strongly correlated with generic quality of life measures (0.47-0.57). In another study non-significant and weak correlations were seen between EQ5D and Quality of Life Scale (QLS) (p N/R). In a third study correlations between EQ5D descriptive system, health states and SQoL-18 dimensions ranged between 0.06 (SQoL family relationship)

						to 0.56 (SQoL self-esteem). Overall correlation with S-QOL index was moderate and significant: 0.48 P<0.05. In one study EQ5D correlated for most part moderately to strongly with WHOQoL-Brief dimension (0.25-0.58) but EQVAS correlations were greater.
				Symptoms and severity	2	EQ5D demonstrated moderate to strong association with one symptom (Clinical Global Impression Severity Scale) and one functional measure (Global Assessment of Function); range 0.34-0.54 p<0.001. In one study weak to moderate correlations were found between EQ5D, Positive and negative symptom scale and Clinical Global Impression Severity Scale (0.189-0.393) (p N/R).
			EQ5D, SF6D	Symptoms and severity	1	In one study correlations between EQ5D and Brief Psychiatric Rating Scale were moderate at baseline (0.343). Also SF6D correlations were moderate but negative (-0.344). Correlations between measures after treatment were weak both for the EQ5D (0.29) and the SF6D (-0.22, p N/A). The two instruments appeared to be responsive only when changes in Brief

						Psychiatric Rating Scale were greater than 25%.
Peasgood	Anxiety/Depression	Good	EQ5D	Symptoms and severity	2	In one study EQ5D correlated at -0.77 with Hamilton Rating Scale for Depression, while in another study the same correlation was found but positive (p N/R).
				HRQoL	3	In one study EQ5D correlated moderately with WHO-BREF (0.545) and EQVAS (0.440). In another study it very strongly correlated (0.7) with the physical health WHO-QoL and strongly (0.5) with mental health WHO-QoL. Correlation was 0.58 overall (p N/R). In one study correlations between SF36 mental health component and EQ5D were moderate to strong (p N/R).
			EQ5D, SF6D	Symptoms and severity	1	In one study EQ5D correlated -0.451 with PHQ-9 Depression test questionnaire at baseline and -0.638 at 3 months follow up. SF6D correlated -0.351 at baseline and -0.833 at 3 months follow up

			SF6D; HUI 3	Symptoms and severity	1	In one study, HUI 3 correlated with Hamilton Anxiety Depression Scale with r= -0.54 and SF6D with r= -0.52; Correlation with Generalized Anxiety Disorder Questionnaire for HUI 3 was -0.44, while for SF6D was -0.38; Correlations with PhQ for HUI 3 was -0.57 and for SF6D was -0.64 (p N/R).
Castelino	Systemic lupus erythematosus	Poor	EQ5D, SF6D	HRQoL	1	In one study, EQ5D convergent validity was -0.69 to -0.55 with corresponding domains of SF36; SF6D had stronger correlations with the SF36 (0.76-0.57; p values N/R). The correlation for the physical component score (0.72; p value N/R) was higher than for the mental component score (0.30; p value N/R).
Davis and Wailoo	Urinary Incontinence	Good	EQ5D	HRQoL	3	In one study statistically significant correlation ($p<0.01$) in the expected direction with the Incontinence Specific Quality of Life Questionnaire were found. In another study correlations between EQ5D and International Continence Society - Benign Prostatic Hyperplasia study Quality of Life Instrument were in the expected direction, some statistically significant and some other not. In a third study King's Health Questionnaire, Patient Generated

						Index correlated in the expected direction with EQ5D (p N/R)
			EQ5D	Symptoms and severity	1	In one study EQ5D correlated strongly (and statistically significantly) with the number of micturition and leakages.
Derrett	Injuries	Poor	EQ5D	HRQoL	3	In one study it was found that EQ5D did not correlate strongly to the EQVAS. In another study the 4 dependent variables “life in general” of LiSat-11, the “general health” scale of SF-36 and the “EQ-5D index” were significantly correlated according to univariate analysis with symptoms and the Coping Strategies Questionnaire. In a third study EQ-5D index correlated with Nottingham Health Profile (0.32, p<0.05), Nottingham pain with EQ-5D pain and disability (0.43, p<0.01) and Nottingham emotional reaction with EQ5D anxiety and depression (0.34, p<0.05).
				Symptoms and severity	2	One study found a weak and statistically significant association between the anxiety and depression domain of EQ5D and the modified

						Dysexecutive Questionnaire (0.29, p<0.05). The remaining four dimensions correlated with the Barthel Index of Activities of Daily Living. In another study the Rivermead Post-concussion symptoms Questionnaire correlated with EQ-5D (r=0.637, p<0.001).
Hounsome	Dementia	Poor	EQ5D	HRQoL	4	Three studies assessed correlations of EQ5D and QoL-AD. In two no association was found between instruments (p<0.0001 in one and N/R in another), in another EQ5D reported by patients correlated strongly with QoL-AD scores (r=0.54; p<0.001). When asked for relevant dimension for their health, participants indicated more HR-QoL attributes than included in EQ5D (e.g. boredom, loneliness, loss of role). In a third study EQ5D scores correlated with those of QoL-AD (r=0.72, p<0.01) and Dementia QoL (r=0.63, p<0.01)
				Symptoms and severity	1	In one study no correlations were found between the mini mental state examination and the EQ5D scores for mobility, pain/discomfort and anxiety/depression (p N/R).
			EQ5D,	HRQoL	1	The EQ5D and HUI 3 correlated with scores

			HUI 3			obtained using Quality of Wellbeing and VAS (p N/R).
			EQ5D	Functional status, symptoms and severity	1	In one study data provided by clinicians had higher construct validity for more observable EQ5D domains (correlation between usual activity domain and Bristol activity of daily leaving scale was very strong (r=0.87, p<0.01). Data provided by carers had higher construct validity for less observable domains (correlation between anxiety/depression and Neuropsychiatric inventory (r=0.57, p<0.01).
Janssen	Type 2 Diabetes Mellitus	Good	EQ5D	HRQoL	5	In one study patients indicating an health problem on the EQ5D had significantly lower mean SF12 component scores for all dimensions. As hypothesized, relationships were stronger between the functional dimension of the SF12 and the physical dimension of the EQ5D and between the mental component of the SF12 and the anxiety depression component of the EQ5D (p N/R). In another study, EQ5D index scores were positively correlated with the summary scores of SF12 (r=0.66 for the Physical component of the SF12, p<0.01 and r=0.41 for the mental component 12, p<0.01). In a third

						study Spearman's correlation between ADDQoL and EQ5D was found to be 0.54, p<0.01. Other two studies found that global self rated status was a predictor of EQ5D scores (p N/R).
				Functional status, symptoms and severity	4	In one study functional status from VFQ-25 and visual acuity (LogMAR) values were independent significant predictors of EQ5D index value. A significant relationship was found between visual acuity and EQ5D index (r^2=0.123). In another, correlation between the weekly mean daily sleep interference score and the EQ5D and VAS were weak to strong (0.13-0.53). In a third study, EQ5D index scores decreased significantly with increasing symptom severity according to the NTSSA-6-SA categories using the Kruskal-Wallis H test. In a fourth one, EQ5D scores decreased significantly according to three severity categories of the Modified Brief Pain Inventory Short Form (p N/R).
			EQ5D, 15 D	HRQoL	1	In one study Spearman's correlations between EQ5D and 15D were 0.81 (Tobit model) and

						0.77 (CLAD model) (p N/R).
Kuspinar and Mayo	Multiple Sclerosis	Excellent	EQ5D, SF6D, HUI 3	HRQoL, functional status	1	In one study the correlation between the EQ5D and the illness instructiveness rating scale was 0.56 (CI 0.45-0.65), whereas SF6D correlation with the same instrument was 0.7 (0.62, 0.77). HUI 3 correlations with the 9-hole pgt test was 0.56 (CI 0.56-0.73) with the Timed 25 foot walk 0.73 (CI 0.66-0.79) with the Ambulation index 0.76 (0.69-0.81) and with the Expanded Disability Status Scale 0.77 (CI 0.7-0.82) (p N/R).
			EQ5D	HRQoL	2	In one study, correlation between patient reported indices for Multiple sclerosis quality of life and EQ5D was 0.58 (0.54-0.62). In another study, EQ5D correlated with: SF36 Physical function 0.26 (CI-0.05-0.52); SF36 Role Physical 0.42 (CI 0.13-0.64); SF36 Bodily Pain 0.20 (CI -0.11-0.48); SF36 Vitality 0.57 (CI 0.32-0.74); SF36 Social Functioning 0.26 (CI -0.05-0.52); SF36 Role Emotional 0.02 (CI -0.29-0.32); SF36 mental health (two results reported: 0.32 (CI 0.16-0.66); 0.44 (CI 0.16-0.66). In a third study, EQ5D correlations with the SF54 Physical Health Composite correlations were

						0.37 (CI 0.18-0.53) and with the SF54 Mental Health Composite 0.42 (CI 0.24-0.57) (p N/R).
			EQ5D, SF6D	Functional status	1	EQ5D correlations were: 9-hole pgt test 0.58 (CI 0.45-0.65); Timed 25 foot walk 0.63 (CI 0.53-0.71); Ambulation index 0.68 (CI 0.59-0.75); Expanded Disability Status Scale 0.66 (0.57-0.63). SF6D correlations were: 9-hole pgt test 0.41 (CI 0.28-0.52); Timed 25 foot walk 0.49 (CI 0.37-0.59); Ambulation index 0.52 (0.41-0.62); Expanded Disability Status Scale 0.48 (CI 0.36-0.58). P N/R.
			EQ5D	Functional status	1	Correlations between EQ5D and Expanded Disability Status Scale were 0.69 (CI 0.6-0.76) (p N/R).
Tosh	Visual impairment	Good	EQ5D	Functional status, symptoms and severity	8	In one study significant correlation were found between visual acuity and EQ5D, while in another one moderate and statistically significant correlations for the mobility, self-care and anxiety dimensions, alongside the summed index score (actual correlations not reported). Three

						studies did not find significant correlations between visual acuity and EQ5D (p N/R). A sixth study reported that lower visual acuity was associated with higher odds of having any problem with all EQ5D dimensions but anxiety and a seventh one that visual angle is a predictor of EQ5D utility values (p N/R). Another study found that worsening of visual acuity was a significant risk factor for a lower EQ5D value (p N/R).
			EQ5D, SF6D, HUI3	Functional status, symptoms and severity	1	SF6D and HUI III significantly correlated with both VA and CS. EQ5D did not present significant correlations (p N/R).
Yang (a)	Hearing impairment	Good	HUI3	Functional status	2	HUI 3 and AN (test that assess suprasegmental identification) 0.33 ($p<0.05$); HUI3 and NVA; NVA (an open speech recognition test) 0.39 ($p<0.05$). In another study the average pure tone air-conduction was a significant predictors of HUI3 ($p<0.01$)
			EQ5D, HUI 3	HRQoL	1	EQ5D with UK or Dutch tariff and HUI 3 correlation between 0.36-0.41. ICC between measures 0.44-0.51 (p N/R).

			EQ5D, SF6D, HUI 3	HRQoL	1	Moderate to strong correlations were found between HUI 3, EQ5D, SF6D (p N/R).
Yang (b)	Skin condition	Good	EQ5D	HRQoL	6	In one study EQ5D reported strong correlations of over 0.5 with Psoriatic Arthritis Quality of Life Scale and patient global VAS that were statistically significant. In another study, EQ5D moderately to strongly correlated with EQVAS and Dermatology Life Quality Index and in a third one weakly to moderately correlated (0.28 to 0.39, p<0.05) with the same instrument. In a fourth study, EQ5D was strongly correlated with Patient's Satisfaction with Life Scale (0.46 p<0.05) and eight domains of SF36 (correlation coefficient ranged between 0.62 and 0.78, p<0.001). In a fifth study strong correlations were found between EQ5D and EQVAS in patients with eczema (P values n/r) and in a sixth moderate to strong correlations between EQ5D a SF36 (p N/R).
				Functional status	1	HAQ disability index is a significant predictor of EQ5D (coefficient -0.31, p<0.05)

Longworth	Cancer	Good	EQ5D	HRQoL	10	In one study significant correlation were found between EQ5D index and EORTC QLQ items (p N/R). In another three studies moderate to strong correlations between EQ5D and EQ-VAS (p N/R). A fifth study reported a correlation of 0.423 between EQ5D and EROTC as opposed to 0.634 between EQVAS and ERTOC. In a sixth study, EQ5D mobility, usual activities and anxiety dimensions and the University of Washington QoL questionnaire overall score were strongly correlated. Strong correlations were also reported between the questionnaire subscales and specific EQ5D dimensions. In a seventh study statistically significant and moderate correlations between all EQ5D dimensions, ECOG and subscales of FACT-G were registered. In an eight study high correlation between EQ5D, EQVAS and EORTC QLQ-C30 ($p<0.001$) were found. A ninth study found moderate to strong and statistically significant correlations between EQ5D, VAS and SF36 (p N/R) and a tenth mentioned that EQ5D and VAS followed a consistent pattern of results (p N/R).

			HUI3	HRQoL	2	HUI 3 showed moderate to very strong positive correlation with SF36 PCS scores and moderate to strong correlation with SF36 MCS. Most correlations were statistically significant. In another study low or non significant correlations were reported between HUI3 and VAS (p N/R) while a third one found moderate correlation between VAS, HUI 3 dimensions and HUI3 index (some statistically significant, some not).
			EQ5D	Functional status	4	In one study, EQ5D and EQVAS were not significantly correlated to the cancer tumour node metastasis stage and the correlation coefficient was low, whereas other measures had moderate correlations. In another study EQ5D index was significantly correlated with the functional assessment of cancer therapy general scale and functional assessment of cancer therapy - kidney symptom index, at 0.6 or above (p N/R). EQ5D and EQVAS were more highly correlated with the condition specific instruments than with each other. Another study found that none of the EQ5D data were significant predictors of Functional Assessment of Cancer Therapy (p N/R). In a fourth study moderate to strong correlation (0.39-0.6) were

						found between EQ5D and Musk skeletal Tumour Rating Scale (p N/R).
			EQ5D, SF6D	HRQoL	1	In one study EORTC subscales (physical and emotional functioning and global health status) were significant predictors of EQ5D, while six EORTC subscales (social and emotional functioning, pain, constipation, dyspnoea and global health status) were significant predictors of SF6D. p values N/R.
			EQ5D, HUI 3	HRQoL	1	Low to moderate correlation between EQ5D and HUI 3 (p N/R).
			HUI 3	Functional status	1	The majority of correlations were moderate to strong between HUI3, Karnofski Performance Status and mini mental state examination (>0.35, p N/R).

Pickard (a)	COPD and Asthma	Good	EQ5D	HRQoL	5	In one study, high correlations were found with Asthma quality of life questionnaire activity, moderate correlations with EORTC QLQ, Asthma quality of life questionnaire symptoms, Asthma quality of life questionnaire emotion and SF-12 PCS, and low correlations with Asthma quality of life questionnaire environment and SF12 MCS (p N/R). In another study changes in EQ5D index significantly correlated with changes in symptoms, activities and emotions of the Asthma Quality of Life Questionnaire (p N/R). In a third study EQ5D showed high correlation with the SF 36 PCS and the St George's Respiratory Questionnaire and moderate correlations with the SF MCS (p N/R). In a fourth study EQ5D was found to highly correlate with SF36 PCS, EQVAS and St George's respiratory Questionnaire (p N/R). In a fifth study, moderate correlations were found with SF36 PCS and MCS and low correlations with Karnofski performance status scale (p N/R).
			EQ5D	Symptoms, severity and functional	2	In one study regression predicting EQ5D in asthma population found post bronco dilatator FEV 1 to be correlated to the index score at p=0.035. In another study mixed performances

				status		of the EQ5D were reported for different functional status measures (p N/R).
Pickard (b)	Cancer	Poor	EQ5D	HRQoL	1	Correlations between VAS and EQ5D were high but actual correlation coefficient was not reported (p N/R).
Sanghera	Menorrhagia	Poor	EQ5D	HRQoL	1	Menorrhagia multi attribute scale was statistically significantly associated with satisfaction of treatment scale whereas EQ5D was not.
Szende	Hemophilia	Good	EQ5D	HRQoL	1	Stronger correlations were reported between SF36 PCS and EQ5D index (0.74) than between the SF36 MCS and the EQ5D index (0.33) (p N/R).
Dyer	Cardiovascular diseases	Good	EQ5D	HRQoL	2	In one study, correlations between EQ5D and SF36 subscales were strong and ranged between 0.57-0.74. Correlations between EQ5D and Mac new subscales ranged between 0.69-0.78 ($p<0.01$). In another study strong correlations were shown between similar EQ5D and SF36 domains for pain (0.68), mobility (0.63) and anxiety/depression (0.75). There was a strong correlation (0.76) between the EQ5D index and

						the general health composite scores for SF36. Moderate to strong correlations were detected between EQ5D and Quality of Life after Myocardial Infarction domain scores, ranging between 0.34-0.56 and with a 0.57 correlation for the index (p N/R).
			EQ5D, SF6D	HRQoL	1	Correlation between the utility domains (EQ5D, SF6D) was rather diffuse with no strong correlations (>0.5) and only a few moderate correlations (p N/R).
			EQ5D, SF6D,HUI3	Symptoms, severity and functional status	1	Moderate to strong correlations were detected by EQ5D, SF6D and HUI 3 with Barthel index of daily living, Modified ranking scale and Centre for epidemiological studies - depression scale. The correaltions were weaker between the VAS, HUI 3 and SF6D (p N/R).
			EQ5D	Symptoms and severity	1	Results showed that low Barthel index scores were associated with very low EQ5D scores (BI of 0 corresponds with EQ5D index of -0.25), demonstrating the sensitivity of the EQ5D for changes in health status (p N/R).

Table VIII – Responsiveness main results

Report	Condition	Quality	GPBM	N studies	Responsiveness test	Main results
Papaioannou (a)	Personality disorder	Good	EQ5D	3	Effect size	In one study effect sizes were found to be large for the long day hospital and short inpatient treatment groups (0.9 and 1.21 respectively) moderate for the long inpatient and long outpatient treatment groups (0.67 and 0.74 respectively) and weak for the short day hospital treatment group. In another study they were 0.47 for the outpatient and 0.59 for the inpatient groups (moderate to weak) and strong for the day hospital group (0.85). In a third study they were weak to moderate pre-post treatment (outpatient 0.37, day hospital 0.72, inpatient 0.32). The comparison of outpatient versus inpatient and day hospital versus inpatient presented weak effect sizes (0.16 and 0.18), whilst they were moderate for outpatient versus day hospital (0.71)
Papaioannou (b)	Schizophrenia	Good	EQ5D	1	Effect size	EQ5D recorded large effect sizes (1.13),

						bigger than EQVAS ones (0.98) pre and post treatment for olanzapine patients and moderate to strong effect sizes for other antipsychotics (0.78 to 0.96)
				1	Standardize Response Mean	When improvements on Brief Psychiatric Rating Scale were at least 25%, EQ5D SRM were small (0.39). When deterioration on Brief Psychiatric Rating Scale was at least 25% or improvement on BPRS <25%, EQ5D SRMs were very small (0.17 to 0.05 respectively)
				1	Change over time	In one study, differences in the EQ5D descriptive system were statistically significant for the daily functioning domain (Z=1.79, P>0.05<0.01) and anxiety depression (Z=3.53, P<0.01).
Peasgood	Depression/ Anxiety	Good	EQ5D	5	Difference between treatment arms	In one study mean difference in QALYs gained between the two groups was 0.00045 (95% CI -0.093; 0.084). Difference in improvements in MADRS score was -0.81 (CI -5.6; 4.0) (p N/R). In another study, no significant difference between intervention and control groups

						was found. However, also clinical measure did not register any difference (p N/R). In a third study patients achieving remission at 8 weeks (SDS<5) reported an EQ5D of 0.87, while patients not achieving remission at 8 weeks (SDS>5) an EQ5D of 0.61. In a fourth study no significant differences were reported between the three treatment groups (p N/R). In a fifth study EQ5D did manage to detect differences between treatment arms only for some groups, whereas clinical instruments detected them for all the groups (p N/R).
				7	Change over time	In one study, EQ5D reported changes from 0.40 (SD 0.01) to 0.73 (SD 0.23) baseline to 6 months follow up (p N/R). EQVAS registered a smaller change. In another study, EQ5D increased from a mean of 0.44 at baseline to a mean of 0.91 at 6 weeks follow up (p N/R). In a third study EQ5D improved from 0.52 to 0.78 baseline to 8 weeks (p<0.001) for the Escitalopram arm and from 0.54 to 0.77 (p<0.001) for the Venlafaxine arm. In a fourth study

						EQ5D showed improvements at 3 and 6 months (p values N/R). In a fifth EQ5D was 0.68 (+- 0.24) at baseline and 0.78 (+- 0.21) at 8 weeks. Extreme difficulty on anxiety and depression were reported by 77.9% of responders at baseline and only from 9.3% of responders at follow up among remitters. In a sixth study EQ5D showed improvements at 3 and 6 months (p values N/R). In a seventh study EQ5D reported substantial improvements from baseline to follow up
				3	Effect size	One study found that UK tariff EQ5D effect sizes for patients were 0.55 while for clinician were 0.65. EQ5D German effect sizes for patients were 0.41 while clinician based effect sizes were 0.45. The Clinical Global Impression scale was found to be the most responsive instrument (both patient based and clinician based). VAS was more responsive than EQ5D but less than CGI. In another study EQ5D was the most responsive instrument with effect size of -0.99 for more anxiety (more than twice that for other measures) and effect

						size of +0.39 for less anxiety. In a third study ES were between 0.31 (0.03-0.6) to 0.67 (0.32-1.02).
			EQ5D, SF6D	1	Standardize Response Mean	The SD of differences were low for SF6D resulting in SRM that were at least twice as high as those for EQ5D for all severity groups. At 1.5 years follow up SRM was 0.466 for EQ5D and 0.833 for SF6D.
				1	Change over time	EQ5D increased 0.147 (change in median score 0.069); SF6D increased 0.082 (change in median score 0.07). They were both statistically significant.
			SF6D	1	Difference between treatment arms	Depression free days and Quality of Well Being found no significant differences between groups, whereas SF6D showed significant differences between intervention and usual care
Castelino	Systemic Lupus Erythematosus	Poor	EQ5D, SF6D	1	Effect size	Both EQ5D (0.012-0.428) and SF6D (0.04-0.43) reported generally small effect sizes.
Ching	Aestetic surgery	Poor	EQ5D	1	Change over	EQ5D seemed to be sensitive to change,

					time	although less sensitive than the SF36
Davis and Wailoo	Urinary Incontinence	Good	EQ5D	3	Change over time	In one study, none of the instruments (EQ5D or clinical instruments) reported a change. In another study significant and statistical improvements (p value N/R) were reported across all instruments in both arms. In a third study, statistically significant improvements in the same direction as other instruments were registered ($p<0.001$).
				1	Standardize Response Mean	EQ5D standardized response mean was 0.26. Symptom severity index SRM was 0.67, Incontinence specific quality of life questionnaire SRM was 1.17 and Incontinence specific quality of life questionnaire domains SRM was 0.8-1.25
				4	Difference between treatment arms	In one study all instruments did not show significant difference between treatment arms and results were statistically significant. In another study, EQ5D registered agreement with clinical outcomes but did not detect differences between arms that were detected by other

						measures. In a third study comparing laser versus resection therapy, only one outcome (maximum flow) presented significant differences in favour of resection. Also EQ5D reported greater gains for resection than laser therapy (p value N/A). In a fourth study, while the number of leaks avoided did not show any significant reduction for the three comparisons between active treatment arms, EQ5D showed statistically significant gains for 2 of three comparisons ($p<0.05$)
Devine	Chronic Low Back Pain	Poor	EQ5D	Meta-analysis	Effect size	The pooled mean effect size for the EQ5D was 0.78 +- 0.12.
Haywood	Older population	Poor	EQ5D	1	Difference between treatment arms	One study reported that EQ5D had the ability to discriminate over treatment groups (p N/R)
Hill	Spinal Cord Injury	Poor	SF6D	1	Difference between treatment arms	In one study it was reported that SF6D could discriminate between the two groups (p values N/R).
Janssen	Diabetes type 2	Good	EQ5D	4	Change over time	In one study a significant decline in health over a 2 years period was found for the EQ5D index (f=5.97, p=0.003) and the

						VAS (f=4.49, p=0.012) as expected. Similar results were found for the EQ5D in another study. In a third study a significant baseline to endpoint change was reported for the insulin group (p=0.049), but not for the Exenatide group (p=0.08). In a fourth study a significant baseline to endpoint change was reported for the more intensive group (p<0.05). For the less intensive self-monitoring and standard usual care group, no significant change was found.
				1	Difference between treatment arms	In one study no statistically significant differences were observed in the routine care or the Duloxine groups (p N/R).
Linder	Acute Synusithis	Excellent	EQ5D	1	Change over time	Change in score (p N/R) were not statistically significant after administering Granulocyte colony-stimulating factor.
Tosh	Visual disorder	Good	HUI3	1	Difference between treatment arms	Statistically significant and appropriate difference between treatment arms (p N/R).
			EQ5D	3	Change over time	In one study, statistically significant improvements were found in both the visual function instrument 4 dimensions

						and the EQ5D after photodynamic therapy in patients with age related macular degeneration. In another study, statistically significant improvements were found in both EQ5D and visual function 14 dimensions post cataract surgery, although EQ5D changes in scores were relatively small. In a third study, where substantial improvements were found by the visual function assessment and visual acuity post cataract surgery, only small and not statistically significant changes were detected by the EQ5D.
Yang (a)	Hearing impairment	Good	EQ5D	2	Change over time	In two studies, EQ5D changes were not significant before and after intervention whereas statistically significant and substantial improvements detected by the two disease specific measures (p some statistically significant)
			EQ5D, HUI3	1	Change over time	In one study, the EQ5D (0.26) and the HUI3 (0.36) increased in score following cochlear implantation, similarly to VAS (0.33) and Quality of Wellbeing (0.16). All

						results were statistically significant.
			HUI 3	3	Change over time	In one study, HUI3 successfully discriminated between people with hearing aids and without hearing aids at 6 months and 12 month after intervention (6 months statistically significant p<0.001; 12 months p<0.1). In another study, HUI3 changes before and after treatments were greater than VAS changes. In a third study, HUI3 seemed to be able to detect a change in score in both group with statistical power, which was in line with those of other instruments.
			EQ5D, HUI3	1	Effect sizes	HUI3 effect size were mdoerate (0.64) whereas EQ5D effect size were small (0.02 UK tariff, 0.05 Dutch tariff).
Yang (b)	Skin condition	Good	EQ5D	5	Difference between treatment arms	In one study, EQ5D values differed between treatment groups (p<0.05) and EQVAS and Disability Life Quality Index confirmed this. In another study standardized mean difference between groups were lower for the EQ5D than for the other instruments (p statistically

						significant). In other three studies the groups that received active treatment achieved significant improvements compared to the placebo groups and this was registered also by other measures, with statistically significant results
				4	Change over time	In one study, EQ5D improved significantly ($p<0.05$) for both treatment arms from baseline to follow up, similarly to the other measures. In another study EQ5D improved significantly (0.17, $p<0.05$) for both treatment arms from baseline to follow up, similarly to the other measures for joint pain. For psoriasis EQ5D did not detected changes that were registered by other measures. In a third study, EQ5D was found to improve significantly (11.5%, $p<0.05$) after 2 weeks, more than EQVAS but less than disease specific instruments. A fourth study found a decrement in the EQ5D and SF36 that had been not reported by the VAS or the short form McGill pain questionnaire (P N/R).

Longworth	Cancer	Good	EQ5D	4	Effect size	In one study large effect sizes were reported for all measures except for the EQ5D. In another study, small effect sizes were reported. A third study reported that EQ5D effect size were comparable to those of other instruments. A fourth study that EQ5D and EORTC had comparable magnitude of change in their effect sizes.
				17	Change over time	In one study EQ5D, Functional Living Index - Cancer and VAS showed a similar patter of change after high dose chemotherapy, and the Friedman test showed significant change over time on four of the five EQ5D dimensions (except pain/discomfort). Results were statistically significant. In another study the EQ5D did not registered any change over time and this was not consistent with the EORTC QLQ-C30 and the Hospital Anxiety and Depression Scale (p N/R). In a third study, significant changes were detected by both faecal incontinence severity and EQ-VAS but not by the EQ5D at 6 months after surgery (p N/R). Another two studies showed that EQ5D followed the same

						patter as the one of other instruments, but p values were N/A. In a sixth study EQ5D followed the same patter followed by the other instruments, but results were not statistically significant. In a seventh study both the EQ5D and the EORTC QLQ-C30 did not detect any difference at follow up (p N/R). One study found EQ5D anxiety/depression and pain dimensions to be statistically significant after treatment. No difference was found for mobility and self-care while statistically significant differences were found on all other outcome measures. Another study reported that improvement registered by the EQ5D were statistically significant at 6 months. One study reported that all EQ5D dimension except for pain/discomfort and EQVAS improved following radiotherapy, but that differences were statistically significant only for the high-risk patients on the EQ5D. Another study reported statistically significant improvements on the anxiety and depression dimensions of EQ5D, in line with the first two concerns

						of Measure Yourself Concerns and Well Being Questionnaire and the overall profile. However, improvements were statistically significant only for the anxiety and depression dimension. These results are similar to those of another study that found only the anxiety, depression and pain dimensions of EQ5D to be statistically significant, whereas the EQVAS was significant in all its dimensions. Other five studies report EQ5D to be responsive (two statistically significant, one no, one at some follow ups, one N/R).
			HUI 3	1	Change over time	Significant decrease in HUI 3 scores shortly after surgery and improvements in the long term were found and were consistent with VAS and SF36. Results were statistically significant.
			EQ5D, HUI3	2	Difference between treatment arms	In one study, the HUI 3 and EQVAS improved in one treatment group but decreased in the other, while EQ5D showed improvements for both arms. Difference between groups were

						statistically significant for HUI3 but not for EQ5D. In another study, EQ5D and HUI3 were less responsive than other measures (p values mixed).
			HUI3	1	Change over time	One study found significant decreases in HUI 3 shortly after surgery and improvements in the long term, consistent with VAS and SF36, that were statistically significant. The emotion attribute of HUI3 was the only dimension that reached statistical significance.
			EQ5D	13	Difference between treatment arms	In three studies EQ5D was able to differentiate between treatment groups. Among them, in one p values were statistically significant, while two did not report them. In a fourth study EQ5D did not differentiated between treatment groups whereas EORTC registered changes (p N/R). Another study reported responsiveness that was consistent with EORCT QLQ-C30 but not EORCT QLQ-C38 (p N/R). A sixth study showed that EQ5D could differentiate between some treatment groups but not all (statistically

						significant).A seventh study reported that EQ5D was able to distinguish between treatment groups (statistically significant differences) similarly to the Rotterdam Symptoms Checklist. In other four studies the EQ5D followed the same patter as the one of other instruments. Results were not statistically significant in three and statistically significant in one. In one study, both the EQ5D and EORTC QLQ-C30 did not manage to differentiate between treatment groups. Another study reported that EQ5D did not discriminated between treatment and control groups. This was consistent with HADS-anxiety and HADS depression but not with the Multidimensional Health Locus of Control Scale chance dimension. Another study reported EQ5D results to be in line with those of EORTC (p N/R).
			HUI 3	1	Difference between treatment arms	In one study, HUI 3 was able to differentiate between treatment groups (p N/R).

			HUI 3	1	Effect size	One study found large and clinically relevant effect sizes, consistent with other measures at two follow ups, but not at the other two time points. Results were statistically significant.
			EQ5D	1	Standardize Response Mean	In the subgroup of patients with no changed global health, neither EQ5D nor EQVAS SRM indicated an effect. For subgroups with a small deterioration or improvement, SRMs of EQ5D were too small to be considered an effect, while SRM of EQVAS showed a small effect. For subgroups with moderate or large improvements or deterioration, SRMs indicated a moderate effect (>0.5) on the EQ5D index and a large effect (>0.8) on the EQVAS.
Petrillo	Asthma/COPD	Poor	EQ5D	3	Change over time	Three studies reported changes over time concordant with expectations. P values were statistically significant in one and N/R in two.
Pickard (a)	Asthma/COPD	Good	EQ5D	2	Standardize Response Mean	One study found a standardized response mean of 0.28 using the EQ5D, while the

						EQVAS SRM was substantially greater. Another study reported a standardized response mean of -0.29 baseline to 6 months.
				2	Effect size	One study found effect size of 0.41 baseline to 3 months while another one effect sizes that were bigger for patients whose global rating improved (ES=-0.55) than for patients with worse global rating improvements (ES=-0.07).
Pickard (b)	Cancer	Good	EQ5D	1	Difference between treatment arms	In one study the EQ5D value changed from 0.73 before needle biopsy to 0.71 after needle biopsy, and from 0.69 before open breast biopsy to 0.61 after open breast biopsy (p N/R).
				1	Change over time	In one study results were in line with the findings of other instruments (p N/R).
Sanghera	Menorrhagia	Poor	EQ5D	1	Change over time	EQ5D lacked sensitivity in changes of QoL reported by other measures after treatment (p N/R).
Wu	HIV	Good	EQ5D	2	Change over	In one study, changes in patient assessed EQ5D scores from baseline to 24 week

					time	follow up were -6 in control group and +8 in switch group with results not being statistically significant (p=0.074), while changes in physician assessed EQ5D scores were -7 in control group and +8 in intervention group, statistically significant (p=0.016) (difficult to understand). In another study, EQ5D scores were significantly higher among patients with improvements in CD4 counts compared to those with no improvements (p<=0.05) with results being similar to MOS-HIV ones.
				3	Effect size	In one study effect size for patients reporting an adverse event were weak (0.4). Instrument dimensions had weak or insignificant (0.05-0.20) effect sizes. In another study, effect size was moderate (0.5) for the usual activities domain (p values not reported) while they were weak for the other dimensions (0.2-0.4). A third study found negligible effect size (0.03-0.05).

Dyer	Heart Disease	Good	EQ5D	3	Effect size	In one study, EQ5D effect sizes for the improved health state patients were 0.74 for the period 1 (not specified) and 0.09 for the period 2 (not specified). VAS had higher effect sizes for the same time points. In another study EQ5D effect sizes were (0.8) and they were greater than for VAS (0.5<). A third study reported that EQ5D index and VAS were less responsive than the Kansas City Cardiomyopathy Questionnaire but showed similar responsiveness to the SF12 PCS and MCS.
				20	Change over time	Six studies reported a change in EQ5D that sustained responsiveness, while four did not detect any change (p N/R). One study reported improvements for both treatment groups from baseline to follow up. However, improvements for the second group were extremely small while VAS improvements were substantial (p N/R). One study reported a similar trend between EQ5D and VAS, although was continued to increase at follow-ups after the 6 months, while EQ5D decreased (p N/R).

						One study registered a decrease in the value of the EQ5D at 5 years follow up not registered by VAS and another study an increase not registered by the same instrument. Other six studies reported a tendency in the expected direction, while another study results not in line with expectations. Three study reported inconsistencies compared to VAS (p N/R).
				2	Difference between treatment arms	Both studies registered differences between treatment arms that were consistent with expectations (p N/R).

Appendices Chapter 4

Table I – Items wordings

	Items	Wording of the first level of the item
EQ-5D-5L	EQ-5D mobility	I have no problems in walking about
	EQ-5D self care	I have no problems washing or dressing myself
	EQ-5D usual activities	I have no problems doing my usual activities
	EQ-5D pain/discomfort	I have no pain or discomfort
	EQ-5D anxiety/depression	I am not anxious or depressed
SF-6D	SF-6D physical functioning	My health does not limit me in vigorous activities
	SF-6D role	I have no problem with my work or other regular daily activities as a result of my physical health or any emotional problem
	SF-6D social functioning	My health limits my social activities none of the time
	SF-6D pain	I have no pain
	SF-6D mental health	I feel tense or downhearted and low none of the time
	SF-6D vitality	I have a lot of energy all of the time
HUI 3	HUI 3 vision	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glass
	HUI 3 hearing	Able to hear what is said in a group conversation with at least three other people, without a hearing aid
	HUI 3 speech	Able to be understood completely when speaking with strangers or people who know me well
	HUI 3 ambulation	Able to walk around the neighbourhood without difficulty, and without walking equipment

	HUI 3 dexterity	Full use of two hands and ten fingers
	HUI 3 emotion	Happy and interested in life
	HUI 3 cognition	Able to remember most things, think clearly and solve day to day problems
	HUI 3 pain	Free of pain and discomfort
AQoL 8D	AQoL 8D energy	[Thinking about how much energy you have to do the things you want to do: I am] Always full of energy
	AQoL 8D social exclusion	[How often do you feel socially excluded or left out?] Never
	AQoL 8D getting around outside house	[Thinking about how easy or difficult it is for you to get around by yourself outside your house (e.g., shopping, visiting)] Getting around is enjoyable and easy
	AQoL 8D role in the community	[Thinking about your health and your role in your community (that is to say neighbourhood, sporting, work, study, church or cultural groups)] My role in the community is unaffected by my health
	AQoL 8D frequency of sadness	[How often do you feel sad?] Never
	AQoL 8D frequency of serious pain	[Thinking about how often you experience serious pain: I experience it] Very rarely
	AQoL 8D confidence in yourself	[How much confidence do you have in yourself?] Completely confident
	AQoL 8D calmness or agitation	When you think about whether you are calm and tranquil or agitated, are you: [Always calm and tranquil]
	AQoL 8D role in the family	[Thinking about your health and your relationship with your family:] My role in the family is unaffected by my health
	AQoL 8D close relationships (family and friends)	[Your close relationships (family and friends) are:] Very satisfying

	AQoL 8D communication	[How well can you communicate with others? (e.g., by talking, listening, writing or signing)] I have no trouble speaking to them or understanding what they are saying
	AQoL 8D frequency trouble sleeping	[How often do you have trouble sleeping?] Never
	AQoL 8D frequency worthlessness	[How often do you feel worthless?] Never
	AQoL 8D frequency feeling anger	[How often do you feel angry?] Never
	AQoL 8D mobility	[Thinking about your mobility, including using any aids or equipment such as wheelchairs, frames, sticks:] I am very mobile
	AQoL 8D self harm	[Do you ever feel like hurting yourself?] Never
	AQoL 8D enthusiasm	[How enthusiastic do you feel?] Extremely
	AQoL 8D frequency worry	[How often did you feel worried over the last 7 days?] Never
	AQoL 8D washing yourself	[Thinking about washing yourself, toileting, dressing, eating or looking after your appearance:] These tasks are very easy for me
	AQoL 8D frequency happiness	[How often do you feel happy?] All the time
	AQoL 8D coping with life problems	[How much do you feel you can cope with life's problems?] Completely
	AQoL 8D intensity of pain/discomfort	[How much pain or discomfort do you experience?] None at all
	AQoL 8D enjoyment of close relationships	[How much do you enjoy your close relationships (family and friends)?] Immensely
	AQoL 8D frequency of pain interfering with usual activities	[How often does pain interfere with your usual activities?] Never

	AQoL 8D frequency of pleasure	[How often do you feel pleasure?] Always
	AQoL 8D frequency of feeling a burden to others	[How much of a burden do you feel you are to other people?] Never
	AQoL 8D contentment with life	[How content are you with your life?] Extremely
	AQoL 8D vision	[How is your vision (while using any visual aids you need)?] I have excellent sight
	AQoL 8D feeling in control of life	[How often do you feel in control of your life?] Always
	AQoL 8D need help for house jobs	[How much help do you need with jobs around the house (e.g., preparing food, cleaning the house or gardening)?] I can do all these tasks very quickly and efficiently without any help
	AQoL 8D social isolation	[How often do you feel socially isolated?] Never
	AQoL 8D hearing	[How is your hearing (while using any hearing aids you need)?] I have excellent hearing
	AQoL 8D frequency depression	[How often do you feel depressed?] Never
	AQoL 8D close relationships (including sexual)	[Your close and intimate relationships (including any sexual relationships) make you:] Very happy
	AQoL 8D frequency of despair	[How often did you feel in despair over the last seven days?] Never
15 D	15 D mobility	I am able to walk normally (without difficulty) indoors, outdoors and on stairs
	15D vision	I see normally, i.e. I can read newspapers and TV text without difficulty (with or without glasses)
	15 D hearing	I can hear normally, i.e. normal speech (with or without a hearing aid)
	15 D breathing	I am able to breathe normally, i.e. with no shortness of breath or other breathing difficulty

	15 D sleeping	I am able to sleep normally, i.e. I have no problems with sleeping
	15 D eating	I am able to eat normally, i.e. with no help from others
	15 D speaking	I am able to speak normally, i.e. clearly, audibly and fluently
	15 D elimination	My bladder and bowel work normally and without problems
	15 D usual activity	I am able to perform my usual activities (e.g. employment, studying, housework, free-time activities) without difficulty
	15 D mental function	I am able to think clearly and logically, and my memory functions well
	15 D discomfort and symptoms	I have no physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc
	15 D depression	I do not feel at all sad, melancholic or depressed
	15 D distress	I do not feel at all anxious, stressed or nervous
	15 D vitality	I feel healthy and energetic
	15 D sexual activities	My state of health has no adverse effect on my sexual activity
ICECAP	ICECAP Feeling settled and secure	I am able to feel settled and secure in all areas of my life
	ICECAP Love, friendship and support	I can have a lot of love, friendship, and support
	ICECAP Being independent	I am able to be completely independent
	ICECAP Achievement and progress	I can achieve and progress in all aspects of my life
	ICECAP Enjoyment and pleasure	I can have a lot of enjoyment and pleasure
ONS	ONS satisfied with life	[Overall, how satisfied are you with your life nowadays?] Completely satisfied

	ONS life is worthwhile	[Overall, to what extent do you feel that the things you do in your life are worthwhile?] Completely worthwhile
	ONS happiness yesterday	[Overall, how happy did you feel yesterday?] Completely happy
	ONSl anxiety yesterday	[Overall, how anxious did you feel yesterday?] Completely anxious
PWI	PWI satisfaction with life as a whole	[Thinking about your own life and personal circumstances, how satisfied are you with your life as a whole?] Completely satisfied
	PWI satisfaction standard of living	[How satisfied are you with your standard of living?] Completely satisfied
	PWI satisfaction health	[How satisfied are you with your health?] Completely satisfied
	PWI satisfaction achievement	[How satisfied are you with what you are achieving in life?] Completely satisfied
	PWI satisfaction personal relationships	[How satisfied are you with your personal relationships?] Completely satisfied
	PWI satisfaction safety	[How satisfied are you with how safe you feel?] Completely satisfied
	PWI satisfaction part of the community	[How satisfied are you with feeling part of your community?] Completely satisfied
	PWI satisfaction future security	[How satisfied are you with your future security?] Completely satisfied
	PWI satisfaction spirituality	[How satisfied are you with your spirituality or religion?] Completely satisfied
SWLS	SWLS life close ideal	[In most ways my life is close to my ideal] How content are you with your life. Strongly agree
	SWLS condition life excellent	[The conditions of my life are excellent] How content are you with your life. Strongly agree
	SWLS satisfaction with life	[I am satisfied with my life] How content are you with your life. Strongly agree
	SWLS gotten important	[So far I have gotten the important things I want

	things in life	in life] How content are you with your life. Strongly agree
	SWLS changing life over	[If I could live my life over, I would change almost nothing] How content are you with your life. Strongly agree

Appendices Chapter 5

Table I – B coefficients, statistical significance and standard errors of factors and items for the first test

Factor to which the item is related	Factor / Item tested	VAS dependent variable	
		B coefficients	Standard errors
/	Satisfaction	*-4.323***	*,112*
/	Relationships	*-5.298***	*,235*
/	Hearing	*-1.209***	*,353*
/	Speech / Cognition	*-2.269***	*,287*
/	Vision	*-2,185***	*.257*
/	Energy / Sleep	*-7.648***	*.217*
Satisfaction	PWI satisfaction standard of living 1	*(omitted)*	*(omitted)*
	PWI satisfaction standard of living 2	*-1.010*	*.811*
	PWI satisfaction standard of living 3	*-2.830***	*.751*
	PWI satisfaction standard of living 4	*-6.449***	*.776*
	PWI satisfaction standard of living 5	*-8.040***	*.844*
	PWI satisfaction standard of living 6	*-10.477***	*.870*
	PWI satisfaction standard of living 7	*-9.834***	*.937*
	PWI satisfaction standard of living 8	*-12.078***	*.993*
	PWI satisfaction standard of living 9	*-14.152***	*1.108*
	PWI satisfaction standard of living 10	*-12.871***	*1.402*
	PWI satisfaction standard of living 11	*-11.483***	*1.438*

Satisfaction	PWI satisfaction achievement 1	*(omitted)*	*(omitted)*
	PWI satisfaction achievement 2	*-.102*	*.860*
	PWI satisfaction achievement 3	*-2.648***	*.815*
	PWI satisfaction achievement 4	*-5.257***	*.832*
	PWI satisfaction achievement 5	*-8.485***	*.890*
	PWI satisfaction achievement 6	*-11.226***	*.869*
	PWI satisfaction achievement 7	*-12.363***	*.998*
	PWI satisfaction achievement 8	*-14.630***	*1.058*
	PWI satisfaction achievement 9	*-17.779***	*1.123*
	PWI satisfaction achievement 10	*-18.501***	*1.258*
	PWI satisfaction achievement 11	*-18.922***	*1.330*
Satisfaction	ONS satisfaction with life 1	*(omitted)*	*(omitted)*
	ONS satisfaction with life 2	*-.361*	*.880*
	ONS satisfaction with life 3	*-4.489***	*.845*
	ONS satisfaction with life 4	*-7.720***	*.883*
	ONS satisfaction with life 5	*-10.704***	*.946*
	ONS satisfaction with life 6	*-11.733***	*.949*
	ONS satisfaction with life 7	*-14.496***	*1.050*
	ONS satisfaction with life 8	*-17.586***	*1.081*
	ONS satisfaction with life 9	*-20.457***	*1.095*
	ONS satisfaction with life 10	*-22.250***	*1.217*
	ONS satisfaction with life 11	*-24.863***	*1.236*
Satisfaction	PWI satisfaction with life as a whole 1	*(omitted)*	*(omitted)*
	PWI satisfaction with life as a whole 2	*-1.018*	*.875*
	PWI satisfaction with life as a whole	*-4.575***	*.800*

	3		
	PWI satisfaction with life as a whole 4	*-8.823***	*.829*
	PWI satisfaction with life as a whole 5	*-12.252***	*.911*
	PWI satisfaction with life as a whole 6	*-12.376***	*.916*
	PWI satisfaction with life as a whole 7	*-15.976***	*1.043*
	PWI satisfaction with life as a whole 8	*-19.238***	*1.060*
	PWI satisfaction with life as a whole 9	*-21.300***	*1.149*
	PWI satisfaction with life as a whole 10	*-19.737***	*1.558*
	PWI satisfaction with life as a whole 11	*-22.271***	*1.604*
Satisfaction	SWLS satisfaction with life 1	*(omitted)*	*(omitted)*
	SWLS satisfaction with life 2	*-3.511***	*.680*
	SWLS satisfaction with life 3	*-7.887***	*.721*
	SWLS satisfaction with life 4	*-12.342***	*.819*
	SWLS satisfaction with life 5	*-14.587***	*.828*
	SWLS satisfaction with life 6	*-19.591***	*.891*
	SWLS satisfaction with life 7	*-24.868***	*1.145*
Satisfaction	SWLS condition of life are excellent 1	*(omitted)*	*(omitted)*
	SWLS condition of life are excellent 2	*-3.293***	*.821*
	SWLS condition of life are excellent 3	*-6.843***	*.828*

	SWLS condition of life are excellent 4	*-11.123***	*.884*
	SWLS condition of life are excellent 5	*-14.229***	*.902*
	SWLS condition of life are excellent 6	*-19.105***	*.951*
	SWLS condition of life are excellent 7	*-23.262***	*1.130*
Satisfaction	SWLS life close ideal 1	*(omitted)*	*(omitted)*
	SWLS life close ideal 2	*-2.136*	*.951*
	SWLS life close ideal 3	*-6.733***	*.952*
	SWLS life close ideal 4	*-9.657***	*1.014*
	SWLS life close ideal 5	*-11.826***	*1.031*
	SWLS life close ideal 6	*-17.383***	*1.049*
	SWLS life close ideal 7	*-20.917***	*1.192*
Satisfaction	SWLS gotten important things in life 1	*(omitted)*	*(omitted)*
	SWLS gotten important things in life 2	*-1.909**	*.623*
	SWLS gotten important things in life 3	*-6.409***	*.658*
	SWLS gotten important things in life 4	*-8.394***	*.750*
	SWLS gotten important things in life 5	*-10.542***	*.781*
	SWLS gotten important things in life 6	*-14.878***	*.859*
	SWLS gotten important things in life 7	*-17.962***	*1.064*
Satisfaction	ONS life is worthwhile 1	*(omitted)*	*(omitted)*

	ONS life is worthwhile 2	*-.991*	*.752*
	ONS life is worthwhile 3	*-4.001***	*.729*
	ONS life is worthwhile 4	*-6.368***	*.760*
	ONS life is worthwhile 5	*-10.142***	*.810*
	ONS life is worthwhile 6	*-11.748***	*.808*
	ONS life is worthwhile 7	*-15.392***	*1.057*
	ONS life is worthwhile 8	*-17.329***	*1.129*
	ONS life is worthwhile 9	*-18.589***	*1.127*
	ONS life is worthwhile 10	*-19.099***	*1.234*
	ONS life is worthwhile 11	*-22.434***	*1.403*
Satisfaction	SWLS if I could live life over 1	*(omitted)*	*(omitted)*
	SWLS if I could live life over 2	*-1.143*	*.830*
	SWLS if I could live life over 3	*-3.139***	*.838*
	SWLS if I could live life over 4	*-4.461***	*.885*
	SWLS if I could live life over 5	*-6.846***	*.838*
	SWLS if I could live life over 6	*-8.685***	*.857*
	SWLS if I could live life over 7	*-12.862***	*.898*
Satisfaction	ONS happiness yesterday 1	*(omitted)*	*(omitted)*
	ONS happiness yesterday 2	*-1.693*	*.719*
	ONS happiness yesterday 3	*-4.224***	*.712*
	ONS happiness yesterday 4	*-7.323***	*.749*
	ONS happiness yesterday 5	*-11.050***	*.816*
	ONS happiness yesterday 6	*-12.203***	*.793*
	ONS happiness yesterday 7	*-13.346***	*1.002*
	ONS happiness yesterday 8	*-15.807***	*1.095*
	ONS happiness yesterday 9	*-17.041***	*1.061*

	ONS happiness yesterday 10	*-18.303***	*1.203*
	ONS happiness yesterday 11	*-19.570***	*1.139*
Satisfaction	PWI satisfaction personal relationships 1	*(omitted)*	*(omitted)*
	PWI satisfaction personal relationships 2	*-1.220*	*.611*
	PWI satisfaction personal relationships 3	*-2.475***	*.627*
	PWI satisfaction personal relationships 4	*-4.891***	*.683*
	PWI satisfaction personal relationships 5	*-5.323***	*.793*
	PWI satisfaction personal relationships 6	*-7.147***	*.729*
	PWI satisfaction personal relationships 7	*-8.671***	*.963*
	PWI satisfaction personal relationships 8	*-10.254***	*1.004*
	PWI satisfaction personal relationships 9	*-10.511***	*1.059*
	PWI satisfaction personal relationships 10	*-11.847***	*1.210*
	PWI satisfaction personal relationships 11	*-13.028***	*1.248*
Satisfaction	PWI satisfaction part of the community 1	*(omitted)*	*(omitted)*
	PWI satisfaction part of the community 2	*-1.042***	*.801*
	PWI satisfaction part of the community 3	*-3.284***	*.753*
	PWI satisfaction part of the community 4	*-5.634***	*.783*

	PWI satisfaction part of the community 5	*-7.638***	*.845*
	PWI satisfaction part of the community 6	*-9.274***	*.757*
	PWI satisfaction part of the community 7	*-10.975***	*1.029*
	PWI satisfaction part of the community 8	*-14.776***	*1.112*
	PWI satisfaction part of the community 9	*-14.396***	*1.149*
	PWI satisfaction part of the community 10	*-15.078***	*1.298*
	PWI satisfaction part of the community 11	*-15.551***	*1.380*
Satisfaction	PWI satisfaction future security 1	*(omitted)*	*(omitted)*
	PWI satisfaction future security 2	*-1.059*	*.880*
	PWI satisfaction future security 3	*-2.746***	*.832*
	PWI satisfaction future security 4	*-4.238***	*.849*
	PWI satisfaction future security 5	*-6.783***	*.905*
	PWI satisfaction future security 6	*-8.868***	*.863*
	PWI satisfaction future security 7	*-9.674***	*.977*
	PWI satisfaction future security 8	*-10.433***	*1.013*
	PWI satisfaction future security 9	*-13.005***	*1.060*
	PWI satisfaction future security 10	*-13.540***	*1.138*
	PWI satisfaction future security 11	*-14.227***	*1.153*
Satisfaction	PWI satisfaction spirituality 1	*(omitted)*	*(omitted)*
	PWI satisfaction spirituality 2	*.010*	*.619*
	PWI satisfaction spirituality 3	*-.236*	*.628*
	PWI satisfaction spirituality 4	*-3.311***	*.729*

	PWI satisfaction spirituality 5	*-4.225***	*.875*
	PWI satisfaction spirituality 6	*-5.102***	*.517*
	PWI satisfaction spirituality 7	*-7.263***	*1.160*
	PWI satisfaction spirituality 8	*-6.449***	*1.345*
	PWI satisfaction spirituality 9	*-8.618***	*1.593*
	PWI satisfaction spirituality 10	*-6.562***	*1.738*
	PWI satisfaction spirituality 11	*-10.234***	*1.348*
Satisfaction	PWI satisfaction safety 1	*(omitted)*	*(omitted)*
	PWI satisfaction safety 2	*-1.557*	*.617*
	PWI satisfaction safety 3	*-3.609***	*.618*
	PWI satisfaction safety 4	*-6.792***	*.680*
	PWI satisfaction safety 5	*-7.524***	*.808*
	PWI satisfaction safety 6	*-9.551***	*.761*
	PWI satisfaction safety 7	*-10.360***	*1.053*
	PWI satisfaction safety 8	*-13.351***	*1.156*
	PWI satisfaction safety 9	*-13.584***	*1.209*
	PWI satisfaction safety 10	*-16.936***	*1.587*
	PWI satisfaction safety 11	*-15.295***	*1.574*
Relationships	AQoL enjoyment close relationships 1	*(omitted)*	*(omitted)*
	AQoL enjoyment close relationships 2	*-2.410***	*.411*
	AQoL enjoyment close relationships 3	*-6.705***	*.633*
	AQoL enjoyment close relationships 4	*-9.235***	*.986*
	AQoL enjoyment close relationships 5	*-11.097***	*3.265*

Relationships	ICECAP Love and support 1	*(omitted)*	*(omitted)*
	ICECAP Love and support 2	*-1.468***	*.420*
	ICECAP Love and support 3	*-5.282***	*.604*
	ICECAP Love and support 4	*-7.657***	*1.561*
Relationships	AQoL close relationships (family and friends) 1	*-2.964***	*.425*
	AQoL close relationships (family and friends) 2	*-6.841***	*.619*
	AQoL close relationships (family and friends) 3	*-11.169***	*.885*
	AQoL close relationships (family and friends) 4	*-15.310***	*1.815*
	AQoL close relationships (family and friends) 5	*-7.748***	*2.095*
Relationships	AQoL close relationships (including sexual) 1	*(omitted)*	*(omitted)*
	AQoL close relationships (including sexual) 2	*-2.619***	*.449*
	AQoL close relationships (including sexual) 3	*-7.399***	*.570*
	AQoL close relationships (including sexual) 4	*-7.383***	*.920*
	AQoL close relationships (including sexual) 5	*-11.277***	*1.365*
Hearing	AQoL hearing 1	*(omitted)*	*(omitted)*
	AQoL hearing 2	*-1.936***	*.443*
	AQoL hearing 3	*-2.576***	*.512*
	AQoL hearing 4	*-4.927***	*1.126*
	AQoL hearing 5	*1.584*	*3.571*
	AQoL hearing 1	*8.191*	*5.420*

Hearing	15 D hearing 1	*(omitted)*	*(omitted)*
	15 D hearing 2	*-1.227**	*.490*
	15 D hearing 3	*-1.360*	*.754*
	15 D hearing 4	*-.047*	*2.090*
	15 D hearing 5	*1.965*	*4.522*
Hearing	HUI 3 hearing 1	*(omitted)*	*(omitted)*
	HUI 3 hearing 2	*-.216*	*.654*
	HUI 3 hearing 3	*-.025*	*.943*
	HUI 3 hearing 4	*-1.665*	*1.086*
	HUI 3 hearing 5	*-4.335**	*1.478*
	HUI 3 hearing 6	*9.290*	*5.761*
Speech / Cognition	AQoL communication 1	*(omitted)*	*(omitted)*
	AQoL communication 2	*-.842*	*.606*
	AQoL communication 3	*-1.707*	*1.243*
	AQoL communication 4	*-1.457*	*2.042*
Speech / Cognition	HUI 3 speech 1	*(omitted)*	*(omitted)*
	HUI 3 speech 2	*-1.445*	*.749*
	HUI 3 speech 3	*.804*	*1.299*
	HUI 3 speech 4	*-2.176*	*2.655*
	HUI 3 speech 5	*-6.090*	*6.177*
Speech / Cognition	15 D speech 1	*(omitted)*	*(omitted)*
	15 D speech 2	*-3.441*	*.702*
	15 D speech 3	*.381*	*1.833*
	15 D speech 4	*3.228*	*3.752*
	15 D speech 5	*-9.872*	*8.141*
Speech / Cognition	15 D mental function 1	*(omitted)*	*(omitted)*

	15 D mental function 2	*-3.557***	*.471*
	15 D mental function 3	*-3.438**	*1.215*
	15 D mental function 4	*-5.656*	*2.338*
	15 D mental function 5	*-1.006*	*6.211*
Speech / Cognition	HUI 3 cognition 1	*(omitted)*	*(omitted)*
	HUI 3 cognition 2	*-3.022***	*.674*
	HUI 3 cognition 3	*-3.674***	*.560*
	HUI 3 cognition 4	*-4.940***	*1.056*
	HUI 3 cognition 5	*-7.719***	*1.993*
	HUI 3 cognition 6	*5.787*	*4.269*
Vision	AQoL vision 1	*(omitted)*	*(omitted)*
	AQoL vision 2	*-2.390***	*.494*
	AQoL vision 3	*-4.070***	*.548*
	AQoL vision 4	*-8.383***	*1.513*
	AQoL vision 5	*-33.079**	*11.450*
	AQoL vision 6	*14.274*	*8.131*
Vision	HUI 3 vision 1	*(omitted)*	*(omitted)*
	HUI 3 vision 2	*-1.489***	*.420*
	HUI 3 vision 3	*-2.187*	*1.046*
	HUI 3 vision 4	*-1.321*	*1.049*
	HUI 3 vision 5	*-9.856***	*2.178*
	HUI 3 vision 6	*-2.377*	*6.165*
Vision	15D vision 1	*(omitted)*	*(omitted)*
	15D vision 2	*-2.286***	*.458*
	15D vision 3	*-3.653***	*1.062*
	15D vision 4	*-1.448*	*1.595*

	15D vision 5	*-1.779*	*5.148*
Energy / Sleep	15D vitality 1	*(omitted)*	*(omitted)*
	15D vitality 2	*-9.071***	*.456*
	15D vitality 3	*-17.609***	*.665*
	15D vitality 4	*-22.761***	*.862*
	15D vitality 5	*-29.180***	*1.279*
Energy / Sleep	AQoL energy 1	*(omitted)*	*(omitted)*
	AQoL energy 2	*-5.036***	*.844*
	AQoL energy 3	*-12.392***	*.885*
	AQoL energy 4	*-19.350***	*.947*
	AQoL energy 5	*-25.957***	*1.194*
Energy / Sleep	SF-6D vitality 1	*(omitted)*	*(omitted)*
	SF-6D vitality 2	*-3.228***	*.879*
	SF-6D vitality 3	*-8.685***	*.898*
	SF-6D vitality 4	*-14.132***	*.961*
	SF-6D vitality 5	*-19.493***	*1.014*
Energy / Sleep	AQoL enthusiasm 1	*(omitted)*	*(omitted)*
	AQoL enthusiasm 2	*-1.937**	*.627*
	AQoL enthusiasm 3	*-7.473***	*.662*
	AQoL enthusiasm 4	*-14.735***	*.826*
	AQoL enthusiasm 5	*-19.979***	*1.358*
Energy / Sleep	AQoL sleeping 1	*(omitted)*	*(omitted)*
	AQoL sleeping 2	*-2.119***	*.620*
	AQoL sleeping 3	*-4.776***	*.633*
	AQoL sleeping 4	*-7.722***	*.735*
	AQoL sleeping 5	*-9.646***	*.893*

Energy / Sleep	15 D sleeping 1	*(omitted)*	*(omitted)*
	15 D sleeping 2	*-2.452***	*.461*
	15 D sleeping 3	*-5.180***	*.580*
	15 D sleeping 4	*-7.396***	*.781*
	15 D sleeping 5	*-10.056***	*1.673*

Note: ** $P \leq 0.05$; ** $P \leq 0.01$*

Table II – B coefficients decrements for chronic conditions using the second test for all items

Dummy	PWI standard living	PWI achieve ment	ONS Satisfaction life	PWI life as whole	SWLS life as whole	SWLS condition life	SWLS life close ideal	SWLS gotten important things	ONS things worthwhil e	SWLS live life over	ONS happiness	PWI person al relatio nships	PWI part commun ity
Cancer	0.023	-0.446	-0.841	-0.397	-0.782	-0.943	-0.573	-0.062	-0.408	0.138	-0.627	0.019	-0.282
Asthma	-0.212	-0.254	-0.259	-0.206	-0.252	-0.513	-0.339	-0.195	-0.169	0.065	-0.381	-0.064	-0.059
COPD	0.155	-0.168	-0.537	-0.436	0.155	0.036	0.472	0.725	-0.103	0.176	-0.276	0.2	-0.132
Depression	-0.232	-0.911	-0.598	-0.863	-0.981	-0.883	-0.65	-0.556	-0.802	-0.591	-0.517	-0.314	-0.592
Diabetes	-0.453	-0.81	-0.888	-0.503	-0.9	-0.832	-0.674	-0.616	-0.832	-0.17	-0.393	-0.166	-0.376
Hearing Problems	0.149	-0.167	-0.284	-0.274	-0.228	0.045	0.04	0.119	-0.204	0.155	-0.193	-0.087	-0.199
Arthritis	0.156	-0.111	-0.223	-0.06	-0.329	-0.177	-0.238	-0.251	-0.097	-0.165	0.114	-0.087	-0.008
Heart	-0.153	-0.593	-0.537	-0.322	-0.342	-0.512	-0.316	0.074	-0.372	0.085	-0.25	0.028	-0.18
Stroke	-1.293	-0.904	-2.582	-2.623	-2.29	-3.519	-1.263	-2.258	-0.828	-1.648	-0.734	-0.532	-1.596

Table II – (continued)

Dummy	PWI future security	PWI religion	PWI safety	AQoL enjoyment close relationships	ICECAP love friendship support	AQoL close rel (family friends)	AQoL close rel (sexual)	AQoL hearing	15D hearing	HUI 3 hearing	15D vision	AQoL vision	HUI 3 vision
cancer	-0.102	-0.252	-0.042	0.07	0.107	0.174	-0.066	-0.322	-0.112	-0.062	-0.052	-0.194	-0.023
asthma	-0.094	0.084	-0.021	-0.062	0.008	-0.089	0.073	-0.165	-0.073	-0.007	-0.038	-0.218	-0.043
COPD	-0.138	-0.174	0.326	0.605	0.179	0.213	0.062	-0.441	-0.226	-0.221	0.058	-0.284	-0.043
Depression	-0.256	-0.127	-0.265	-0.696	-0.138	-0.677	-0.537	-0.153	-0.055	-0.037	0.052	-0.098	-0.101
Diabetes	-0.28	-0.078	0.031	-0.111	0.07	-0.118	-0.139	-0.181	-0.089	-0.067	-0.149	-0.323	-0.139
Hearing Problems	0.003	0.034	-0.184	-0.275	-0.074	-0.231	-0.106	-1.663	-0.974	-0.706	-0.208	-0.364	-0.184
Arthritis	0.048	0.051	0.234	-0.013	0.097	-0.02	0.248	-0.182	-0.059	-0.023	-0.004	-0.187	-0.055
Heart	-0.169	0.022	0.032	0.143	0.02	0.027	0.111	-0.305	-0.108	-0.083	-0.057	-0.208	-0.063
Stroke	-1.732	-0.746	-1.201	-0.28	-1.01	0.009	-0.473	-0.65	-0.153	-0.147	-0.477	-0.832	-0.751

Table II – (continued)

Dummy	AQoL communication	HUI 3 speech	15D speech	15D mental functioning	HUI3 cognition	15D vitality	AQoL energy	SF6D vitality	AQoL enthusiasm	AQoL sleep	15D sleep
cancer	-0.027	0.001	-0.019	-0.296	-0.273	-1.733	-1.802	-1.048	-0.383	-0.434	-0.263
asthma	0.002	0.017	-0.059	-0.07	-0.101	-1.709	-1.618	-1.02	-0.377	-0.439	-0.455
COPD	0.019	-0.017	0.148	0.08	0.281	-2.162	-3.059	-2.187	-0.36	-0.451	-0.59
Depression	-0.048	0.005	-0.1	-0.456	-0.44	-2.707	-2.836	-1.975	-1.763	-0.807	-0.649
Diabetes	0.008	0.016	-0.008	-0.136	-0.087	-1.835	-1.914	-1.16	-0.544	-0.459	-0.279
Hearing Problems	-0.224	-0.022	-0.171	-0.327	-0.376	-1.13	-1.194	-0.561	-0.448	-0.296	-0.111
Arthritis	0.02	0.032	-0.017	-0.01	0.013	-0.786	-0.907	-0.427	-0.403	-0.477	-0.34
Heart	-0.003	0.01	-0.057	-0.08	-0.154	-2.031	-2.284	-1.363	-0.635	-0.444	-0.355
Stroke	-0.202	-0.139	-0.916	-1.137	-1.293	-2.959	-2.401	-0.265	-1.136	-0.484	-0.561

Appendices Chapter 6

Material I – Grant approval letter

EQ-5D

Aureliano Paolo Finch
School of Health and Related Research
University of Sheffield
West Court
1 Mappin Street
S1 4DT
Sheffield

31Mar2017

Dear Aureliano,

Thank you for submitting your fast-track research proposal. The Executive Committee of the EuroQol Group has agreed to make available a budget of €9,313 for your research project titled *"Testing the impact of potential bolt-ons on preferences using pairwise choices: A pilot study."* by fast-track procedure on 31MAR2017, as outlined in your project proposal, see appendix I.

As is the case with all proposals funded by the EuroQol Research Foundation ("the Foundation"), the amount you requested for your accepted proposal will be made available subject to the following conditions:

i) Funding of all projects by the Foundation is conditional on adherence to certain principles that underpin the activities of the Foundation. Prominent among those is that copyright of EQ-5D products is dealt with at Foundation level to ensure the future financing of the Foundation.

ii) Any substantive changes or delays to your project must be reported back to the Executive Committee.

iii) Manuscripts must include the disclaimer statement that the views expressed by the authors in the publication do not necessarily reflect the views of the EuroQol Group.

50% (€4,657) of the budget will be transferred to your institute immediately. An invoice, making reference to **EQ Project 20170210**, from your institute should be sent to Bernhard Slaap (please provide bank name, bank address, account number/IBAN and BIC code) at the address below.

Bernhard Slaap
EuroQol Executive Director
Marten Meesweg 107
3068 AV Rotterdam
The Netherlands
slaap@euroqol.org

If preferred, co-workers can invoice me directly for this research project. Note that the principle investigator is responsible for the study budget, so you need to keep track of invoicing. Please use the **EQ project number**, see above, for all invoices.

Template version 15MAR2017

EQ-5D

A report (approximately 1000 words) should be provided to Scientific Team Leader Elly Stolk (stolk@euroqol.org) on completion of the project. For the reporting requirements please visit the Members' Area of our website. The remaining budget (€4,546) will be transferred to your institution after signoff by the Executive Committee of your report.

We look forward to the results of this exciting project.

Please can you make two copies of this document, sign them both and return one copy to me (a signed pdf will suffice). The other should be retained for your files.

Yours sincerely,

Signature:	[redacted] Digitally signed by slaap@euroqol.org Date: 2017.03.31 15:17:31 +02'00'	Signature:	[redacted]
Date:		Date:	31.03.2017
Name:	Bernhard Slaap	Name:	AURELIANO PAOLO FINCH
Title:	Executive Director	Title:	MR
Organization:	EuroQol Research Foundation	Organization:	UNIVERSITY OF SHEFFIELD

Appendixes:

I: Study proposal, as agreed by the EQ Executive Committee

Material II – Ethics approval focus group 1

Downloaded: 05/08/2017
Approved: 10/10/2016

Aureliano Paolo Finch
Registration number: 140102965
School of Health and Related Research
Programme: HARR41 Health and Related Research

Dear Aureliano Paolo

PROJECT TITLE: Testing the face validity of bolt-ons wording
APPLICATION: Reference Number 010317

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 10/10/2016 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 010317 (dated 07/10/2016).
- Participant information sheet 1023236 version 2 (04/10/2016).
- Participant consent form 1023237 version 2 (04/10/2016).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely

Jennifer Burr
Ethics Administrator
School of Health and Related Research

Material III – Ethics approval for focus group 2

The University Of Sheffield.

School Of Health And Related Research.

ScHARR

Ellen Nicolson
Ethics Committee Administrator
Regent Court
30 Regent Street
Sheffield S1 4DA

27 January 2017

Telephone: +44 (0) 114 222 5446
Fax: +44 (0) 114 222 0749 (non confidential)
Email: e.l.nicolson@sheffield.ac.uk

Project title: Testing the face validity of bolt-ons wording

Reference Number: 010317

Dear Aureliano,

Thank you for submitting the above amended research project for approval by the ScHARR Research Ethics Committee. On behalf of the University, I am pleased to inform you that the project with changes 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

[redacted]

Ellen Nicolson
On behalf of the ScHARR Research Ethics Committee

Material IV – Invitation e-mail focus group 1

Invitation email

My name is Aureliano Paolo Finch and I am a PhD student at the University of Sheffield in the School of Health and Related Research. My PhD project investigates methods to identify, develop and select dimensions of health and wellbeing that can be added to an existing measure, to better describe patients' health status in different conditions.

You are kindly invited to participate to a focus group that is part of this broader PhD project. The focus group will ask your opinion on the wording of some items that have been added to an existing measure. A single focus group will be organised in October at a mutually convenient time. Further information about the study is provided below (and as an attachment). If you have any further questions please do not hesitate to get in touch with me.

If you are interested in taking part, please reply to this email with details of your telephone number as well as your preference in terms of days and times (AM/PM) to attend the focus group.

I look forward to hearing from you.

Kind Regards,

Aureliano P Finch

West Court

Mappin Street 2

S1 4DT

Sheffield

E-mail: APFinch1@sheffield.ac.uk

Material V – Invitation e-mail focus group 2

Invitation email

Are you affected by a chronic health condition? Do you have problems such as, but not limited to, asthma, chronic obstructive pulmonary disease, heart diseases, diabetes, epilepsy, multiple sclerosis or arthritis?

If so, I would really appreciate to hear your views on some dimensions of health and wellbeing that have been added to an existing measure.

You are kindly invited to participate to a focus group that will be held **the 13th of February** at a mutually convenient time. Further information about the study is provided below (and as an attachment). If you have any questions please do not hesitate to get in touch with me.

If you are interested in taking part, please reply **to APFinch1@sheffield.ac.uk** with details of your telephone number as well as your preference in terms of times (AM/PM) to attend the focus group.

I look forward to hearing from you.

Kind Regards,

Aureliano P Finch

West Court

Mappin Street 2

S1 4DT

Sheffield

E-mail: APFinch1@sheffield.ac.uk

Material VI - Information sheet focus group 1

Information sheet

Research Project Title

Testing the face validity of additional dimensions on an existing measure.

Invitation

You are being invited to take part in a research project. Before you decide whether to participate, it is important for you to understand the reasons for this research and what your participation will involve. Please read the information reported here carefully, and discuss them with others if you wish. If something is not clear to you, or you wish further details on some aspects of this study, we are happy to answer your questions.

Why is this study being conducted?

Some measures of health are routinely used to assess the impact of new treatments and interventions on the quality of life of patients, with the EQ-5D being the preferred measure in the United Kingdom. Although the EQ-5D is meant to be applicable across all conditions, evidence exists that it might not function appropriately in some situations. In those cases, a possible solution is to add one or more items describing the dimension/s of health missed by the measure. This study tests the wording of a set of potential items identified through previous research.

Why have I been chosen?

You have been invited to participate as you are currently working or studying in the School of Health and Related Research at the University of Sheffield.

Is my participation mandatory?

No, you are completely free to decide whether to take part or not to this study. If you would like to participate, please reply to this email. Even if you agree to participate to the study, you will always be able to withdraw at any time without having to give any reason. No more information will be collected from you if you decide to withdraw from the study after agreeing to participate.

What will I have to do if I decide to take part to the study?

You will be asked to provide dates in which you might be available to participate to the focus group, which will be held in one of the University of Sheffield buildings. On the day of the focus group, you will be given a consent form to sign. You will complete questionnaires with additional dimensions on them for yourself before discussing with the other participants what you think of them. The researcher may ask questions to help the discussion. You will also be asked a few questions about yourself. The focus group will take up to two hours.

Will the focus group be recorded, and how will the recorded media be used if this is the case?

Yes, the whole focus group will be audio recorded (not filmed) and the researcher will take notes during the discussion. All audio recordings will be transcribed and personal references will be removed from the transcript, using pseudonyms. If you decide to withdraw, data will be recorded till the point of your withdrawal. Only the research team will have access to the original recordings, which will be deleted after being transcribed. The completed questionnaires will be retained and your answers will be recoded without any identifying information. All consent forms, transcripts and questionnaires will be stored in a safe and secure location on the campus of the University of Sheffield, where only members of the research team have access.

What are the possible benefits of taking part to this focus group?

There are no direct benefits of taking part but it is anticipated that the results of the focus group will aid in the development of measures which are used in decision-making that has an impact on everyone.

What are the possible disadvantages and risks of taking part?

There are no anticipated negative effects of this study.

What should I do in case something goes wrong?

If something goes wrong, please contact John E Brazier at the telephone number (+44) (0)114 222 0726 or via e-mail at j.e.brazier@sheffield.ac.uk. If this does not resolve the issue or you would prefer to speak to someone else, please contact Jon Nicholl at the telephone number (+44) (0)114 222 5453 or via mail at j.nicholl@sheffield.ac.uk.

Will my participation in this project be kept confidential?

All the information collected in this project will be confidential apart from other study participants. You and your opinions will not be identifiable in any future dissemination activity, including reports or publications. All references to your personal details will be removed from the audio recordings and these will be deleted once transcribed. Transcripts will be anonymised and archived on the university campus.

How will the results of this focus group be used?

The results of this focus group will inform subsequent research projects, which will form part of a PhD thesis. Some of the data collected in the focus group may be used for journal articles or conference presentations. All results that will be made public will be anonymised.

Who is organising and funding the research?

The research is part of a PhD study that is funded by a departmental studentship from the School of Health and Related Research of the University of Sheffield.

Who has ethically reviewed the project?

The School of Health and Health Related Research of the University of Sheffield has reviewed and approved this project.

Contact for further information

Aureliano Paolo Finch

West Court

Mappin Street 2

S1 4DT

Sheffield

E-mail: APFinch1@sheffield.ac.uk

Material VII – Information sheet focus group 2

Information sheet

Research Project Title

Testing the face validity of additional dimensions on an existing measure.

Invitation

You are being invited to take part in a research project. Before you decide whether to participate, it is important for you to understand the reasons for this research and what your participation will involve. Please read the information reported here carefully, and discuss them with others if you wish. If something is not clear to you, or you wish further details on some aspects of this study, we are happy to answer your questions.

Why is this study being conducted?

Some measures of health are routinely used to assess the impact of new treatments and interventions on the quality of life of patients, with the EQ-5D being the preferred measure in the United Kingdom. Although the EQ-5D is meant to be applicable across all conditions, evidence exists that it might not function appropriately in some situations. In those cases, a possible solution is to add one or more items describing the dimension/s of health missed by the measure. This study tests the wording of a set of potential items identified through previous research.

Why have I been chosen?

You have been invited to participate as **you are on the University of Sheffield Volunteer list.**

Is my participation mandatory?

No, you are completely free to decide whether to take part or not to this study. If you would like to participate, please reply to this email. Even if you agree to participate to the study, you will always be able to withdraw at any time without having to give any reason. No more information will be collected from you if you decide to withdraw from the study after agreeing to participate.

What will I have to do if I decide to take part to the study?

You will be asked to provide times in which you might be available to participate to the focus group, which will be held in one of the University of Sheffield buildings. On the day of the focus group, you will be given a consent form to sign. You will complete questionnaires with additional dimensions on them for yourself before discussing with the other participants what you think of them. The researcher may ask questions to help the discussion. You will also be asked a few questions about yourself. The focus group will take up to two hours.

Will the focus group be recorded, and how will the recorded media be used if this is the case?

Yes, the whole focus group will be audio recorded (not filmed) and the researcher will take notes during the discussion. All audio recordings will be transcribed and personal references will be removed from the transcript, using pseudonyms. If you decide to withdraw, data will be recorded till the point of your withdrawal. Only the research team will have access to the original recordings, which will be deleted after being transcribed. The completed questionnaires will be retained and your answers will be recoded without any identifying information. All consent forms, transcripts and questionnaires will be stored in a safe and secure location on the campus of the University of Sheffield, where only members of the research team have access.

What are the possible benefits of taking part to this focus group?

There are no direct benefits of taking part but it is anticipated that the results of the focus group will aid in the development of measures, which are used in decision-making that has an impact on everyone.

What are the possible disadvantages and risks of taking part?

Although no potential harm is anticipated from taking part, the focus group will discuss the content of health measures, which may be areas in which participants have a problem and this may be distressing. Participants are free not to take part or to leave the focus group early if this happens.

What should I do in case something goes wrong?

If something goes wrong, please contact John E Brazier at the telephone number (+44) (0)114 222 0726 or via e-mail at j.e.brazier@sheffield.ac.uk. If this does not resolve the issue or you would prefer to speak to someone else, please contact Jon Nicholl at the telephone number (+44) (0)114 222 5453 or via mail at j.nicholl@sheffield.ac.uk.

Will my participation in this project be kept confidential?

All the information collected in this project will be confidential apart from other study participants. You and your opinions will not be identifiable in any future dissemination activity, including reports or publications. All references to your personal details will be removed from the audio recordings and these will be deleted once transcribed. Transcripts will be anonymised and archived on the university campus.

How will the results of this focus group be used?

The results of this focus group will inform subsequent research projects, which will form part of a PhD thesis. Some of the data collected in the focus group may be used for journal articles or conference presentations. All results that will be made public will be anonymised.

Who is organising and funding the research?

The research is part of a PhD study that is funded by a departmental studentship from the School of Health and Related Research of the University of Sheffield.

Who has ethically reviewed the project?

The School of Health and Health Related Research of the University of Sheffield has reviewed and approved this project.

Contact for further information

Aureliano Paolo Finch

West Court

Mappin Street 2

S1 4DT

Sheffield

E-mail: APFinch1@sheffield.ac.uk

Material VIII - Background questionnaire focus group 1

Background information

Age:

☐ 18-25
equivalent

☐ 25-34

☐ 35-44

☐ 45-54

☐ 55+

☐ First degree

☐ University higher degree

What is the highest education you have achieved?

☐ O-Level/GCSE or

☐ A-Level

☐ Diploma

Gender:

☐ Female

☐ Male

Do you consider yourself religious?

☐ Yes

☐ No

What is your professional activity?

☐ Student

☐ Employed or self employed

☐ Retired

Have you ever used the EQ-5D or any similar types of questionnaires?

☐ Yes

☐ No

☐ Unemployed

What is your marital status?

☐ Married or living with a partner

☐ Single or never married

☐ Divorced, separated or widowed

Do you or someone close to you have any previous experience of serious illness?

☐ No

☐ Myself

☐ Family or friends

☐ Others

Do you have any children?

☐ Yes

☐ No

Material IX – Background questionnaire focus group 2

Background information

Age:

☐ 18-25

☐ 25-34

☐ 35-44

☐ 45-54

☐ 55+

What is the highest education you have achieved?

☐ O-Level/GCSE or equivalent

☐ A-Level

☐ Diploma

☐ First degree

☐ University higher degree

Gender:

☐ Female

☐ Male

Do you consider yourself religious?

☐ Yes

☐ No

What is your professional activity?

☐ Student

☐ Employed or self employed

☐ Retired

Have you ever used the EQ-5D or any similar types of questionnaires?

☐ Yes

☐ No

☐ Unemployed

What is your marital status?

☐ Married or living with a partner

☐ Single or never married

☐ Divorced, separated or widowed

What is your chronic health condition?

____

Do you have any children?

☐ Yes

☐ No

Material X - Topic guide focus group 1 and 2

Focus Group Topic Guide

Testing the face validity of bolt-ons wordings

This focus group will comprise three phases. The first phase will briefly introduce the objective of the study and its main steps. The second phase will familiarise participants with the measure investigated, the EQ-5D, and it will collect preliminary information on its comprehensiveness and easiness of administration. The third phase will explore participants' ideas on the bolt-ons relevance, clarity, easiness of administration (easiness to understand and to respond to) and acceptability in different populations. This topic guide presents the main probes for each of the focus group phases, as well as their expected duration.

First phase

- Participants will be briefly presented the aims and objectives of the focus group and its organisation. They will be reminded that the session is audio recorded and invited to discuss their ideas without any fear of giving a wrong answer.

(Approximately 10 minutes)

Second phase

- Participants will be administered the standard EQ-5D and they will be asked to complete it. Once the measure has been completed they will be asked to discuss their thoughts about it.

  Probes

  1. What do participants think about the measure in terms of its ability to detect changes in quality of life?
  2. Is it easy to complete?

3. Does it cover most of the dimension of health that are relevant and if not, would they add any dimension?

(Approximately 30 minutes)

Third phase

- The set-up of the third phase will mostly depend on the information collected during the second phase. For example, if in the second phase participants claimed that the EQ-5D is missing sensory dimensions, the three sensory bolt-ons (vision, hearing and speech) will be presented as an example and participants will be asked to discuss them. This will be followed by a discussion of the other bolt-ons, organised by theme. An attempt will be made to discuss all bolt-ons, with the wellbeing ones (satisfaction and close relationships) being left as last.

Probes

4. Is/Are the additional dimension/s relevant? i.e. Is the ability to detect changes in quality of life improved with the additional dimension/s? If yes, for all population or just some?
5. Is/Are the additional dimension/s easy to understand? E.g. clarity of the language, unambiguity of levels. If not, how would participant improve it/them?
6. Is/Are the additional dimension/s easy to respond to?
7. What did participants think when completing the additional item/s?
8. Does the measure appear consistent in the way it describes health with the additional dimension/s?

(Approximately 80 minutes)

Materials XI – EQ-5D questionnaire focus groups 1 and 2

EQ-5D questionnaire

MOBILITY

I have no problems in walking about ☐
I have slight problems in walking about ☐
I have moderate problems in walking about ☐
I have severe problems in walking about ☐
I am unable to walk about ☐

SELF-CARE

I have no problems in washing or dressing myself ☐
I have slight problems in washing or dressing myself ☐
I have moderate problems in washing or dressing myself ☐
I have severe problems in washing or dressing myself ☐
I am unable to wash or dress myself ☐

USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activity)

I have no problems doing my usual activities ☐
I have slight problems doing my usual activities ☐
I have moderate problems doing my usual activities ☐
I have severe problems doing my usual activities ☐
I am unable to do my usual activities ☐

PAIN/ DISCOMFORT

I have no pain or discomfort ☐

I have slight pain or discomfort ☐

I have moderate pain or discomfort ☐

I have severe pain or discomfort ☐

I have extreme pain or discomfort ☐

ANXIETY / DEPRESSION

I am not anxious or depressed ☐

I am slightly anxious or depressed ☐

I am moderately anxious or depressed ☐

I am severely anxious or depressed ☐

I am extremely anxious or depressed ☐

Material XII – Invitation email survey

Invitation email

You are kindly invited to participate to a research project. This will take the form of a survey where you will be presented a number of health conditions. You will be asked to imagine having to live in those health conditions till you die and you will be asked to choose which of them is more desirable. Health conditions will be presented in pairs. The results of this project will help to better understand the aspects of health that are important for people and that should be therefore used to describe health in a commonly used measure.

If you are interested in having additional information on this survey or to take part to it, please click the link below.

Link:

Material XIII – Information sheet survey

Information sheet

Welcome to the "Examining aspects of health to expand a measure" project

Thank you for your interest in the "Examining aspects of health to expand a measure" project. Before you decide whether to participate, it is important for you to understand the reasons for this research and what your participation will involve. Please read the information reported here carefully.

Why is this study being conducted?

To decide whether a new drug should be made available to patients, the NHS uses a measure of health, called the EQ-5D. This measure can detect changes in the quality of life of people in most diseases. However, it does not work well in a few others. One of the reasons why this might happen is that this measure does not ask questions on aspect of the quality of life that might be relevant for some people. This study tests the relevance of some additional questions for this measure.

Why have I been chosen?

You have been invited to participate as you are a member of the general public. However, we kindly ask you not to respond to the survey if you are under 18 and/or you cannot read or speak English fluently, as in this case you received this invitation by mistake.

Is my participation mandatory?

No, you are completely free to decide whether to take part or not to this study. To have additional information on this study please continue reading the information below. Please be aware that even if you agree to participate to the study, you will always be able to withdraw at any time without having to give any explanation, and your information will not be used.

What will I have to do if I decide to take part to the study?

You will be asked to consent to participate to the study. Subsequently, you will be presented with a choice between two health conditions. You will be asked to imagine living in those health conditions till you die, and to choose which of them you would prefer. You will also be asked to complete a short questionnaire on your health and some background information.

What are the possible benefits of taking part to this focus group?

There are no direct benefits of taking part in this study. However, you will receive some points that can be exchange for goods once a given number of points is reached.

What are the possible disadvantages and risks of taking part?

Thinking about serious ill states can be upsetting. In addition, in responding to the choices, you will have to imagine living in those ill health states till you die. You may also have to judge what is the value of living your life with certain ill states. However, as this study is interested in understanding your perspective and ideas, there is not a right or wrong answer.

What should I do in case something goes wrong?

If something goes wrong, please contact John E Brazier at the telephone number (+44) (0)114 222 0726 or via e-mail at j.e.brazier@sheffield.ac.uk. If this does not resolve the issue or you would prefer to speak to someone else, please contact Jon Nicholl at the telephone number (+44) (0)114 222 5453 or via mail at j.nicholl@sheffield.ac.uk.

Will my participation in this project be kept confidential?

All the information collected in this project will be confidential. You and your opinions will not be identifiable in any future dissemination activity, including reports or publications. All references to your personal details will be removed once the full sample has been achieved.

How will the results of this survey be used?

The results of this survey will inform subsequent research projects, which will form part of a PhD thesis. Some of the data collected in this survey may be used to publish journal articles or for conference presentations. All results that will be made public will be anonymised.

Who is organising and funding the research?

The research is part of a PhD study that is funded by a departmental studentship from the School of Health and Related Research of the University of Sheffield.

Who has ethically reviewed the project?

The School of Health and Health Related Research of the University of Sheffield has reviewed and approved this project.

Material XIV – Consent form survey

The University Of Sheffield.

Consent form

Title of the research project: Examining aspects of health to expand a measure

PLEASE READ CAREFULLY AND TICK THE FOLLOWING FORM

	Please tick box
1. I have read and understood the information sheet dated *[insert date]*. I understand what the project is about.	☐
2. I understand that my participation is entirely voluntary and that I can decide to withdraw from the survey at any time without having to give any reason.	☐
3. I understand that anything I say will be kept confidential.	☐
4. I give permission to the members of the project team to have access to my anonymised answers.	☐
5. I understand that my name will not be linked with the research material and that I will not be identified or identifiable from the answers I give in any reports that will result from this research.	☐
6. I agree for the anonymised data collected from me to be used in future projects.	☐
7. I am happy to take part in the project.	☐

Material XV – Survey

“Examining aspects of health to expand a measure” project

Please answer the following questions

What is your gender?

Male ☐

Female ☐

What is your date of birth?

Please enter your date of birth in DD/MM/YYYY format

Are you

Married / Partner ☐

Single ☐

Separated ☐

Divorced ☐

Widowed ☐

Prefer not to say ☐

Do you have any children?

Yes

No

Which of the following best describes your main activity?

In employment or self-employment

Retired

Homemaker

Student

Seeking work

Unemployed

Long term sick

Other

What is the highest education you have achieved?

O-level / GCSE or equivalent

A-levels

Diploma

First degree

Postgraduate degree

Do you consider yourself religious?

Yes

No

Prefer not to respond

By placing a tick in one box in each group below, please indicate which statements best describe your own health today

MOBILITY

I have no problems in walking about ☐

I have slight problems in walking about ☐

I have moderate problems in walking about ☐

I have severe problems in walking about ☐

I am unable to walk about ☐

SELF-CARE

I have no problems in washing or dressing myself ☐

I have slight problems in washing or dressing myself ☐

I have moderate problems in washing or dressing myself ☐

I have severe problems in washing or dressing myself ☐

I am unable to wash or dress myself ☐

USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activity)

I have no problems doing my usual activities ☐

I have slight problems doing my usual activities ☐

I have moderate problems doing my usual activities ☐

I have severe problems doing my usual activities ☐

I am unable to do my usual activities ☐

PAIN/ DISCOMFORT

I have no pain or discomfort ☐

I have slight pain or discomfort ☐

I have moderate pain or discomfort ☐

I have severe pain or discomfort ☐

I have extreme pain or discomfort ☐

ANXIETY / DEPRESSION

I am not anxious or depressed ☐

I am slightly anxious or depressed ☐

I am moderately anxious or depressed ☐

I am severely anxious or depressed ☐

I am extremely anxious or depressed ☐

HEARING

I have no problems hearing ☐

I have slight problems hearing ☐

I have moderate problems hearing ☐

I have severe problems hearing ☐

I am unable to hear ☐

SLEEP

I have no problems sleeping ☐

I have slight problems sleeping ☐

I have moderate problems sleeping ☐

I have severe problems sleeping ☐

I have extreme problems sleeping ☐

RELATIONSHIPS

I have no problems with my social relationships ☐

I have slight problems with my social relationships ☐

I have moderate problems with my social relationships ☐

I have severe problems with my social relationships ☐

I am unable to have social relationships ☐

ENERGY

I have no problems with my energy levels ☐

I have slight problems with my energy levels ☐

I have moderate problems with my energy levels ☐

I have severe problems with my energy levels ☐

I have extreme problems with my energy levels ☐

COGNITION

I have no problems with remembering things ☐

I have slight problems with remembering things ☐

I have moderate problems with remembering things ☐

I have severe problems with remembering things ☐

I am unable to remember things ☐

“Examining aspects of health to expand a measure” survey

The next section presents two imaginary descriptions of health (scenarios A and B) based on some of the questions about health from the previous section. Each scenario describes health in a different way.

We want you to imagine what it would be like to live in each of them for 10 years without relief or treatment and then die. Please imagine that death will be very swift and completely painless. Please also imagine that you will have no other problems besides what is indicated.

We then want you to tell us which of the two health scenarios you would prefer.

Please read each description carefully before making your choice as the descriptions vary across the questions.

There are no right or wrong answers to any of the questions.

Each scenario includes a combination of five or six of the following areas of health:

- Mobility
- Self-care
- Usual activities
- Pain or discomfort
- Anxiety or depression
- Hearing
- Sleep
- Relationships
- Energy
- Cognition

Two examples of the sort of questions that you will be asked are below. Please complete the examples.

Example 1

	Health scenario A	*Health scenario B*
	- Severe problems in walking about - Moderate problems in washing or dressing yourself - Severe problems doing your usual activities - Moderate pain or discomfort - Not anxious or depressed	- Severe problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Moderate pain or discomfort - Severely anxious or depressed
Which scenario do you think is better?	☐	☐

Example 2

	Health scenario A	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Not anxious or depressed	- Severe problems in walking about - Severe problems in washing or dressing yourself - Unable to do your usual activities - Moderate pain or discomfort - Severely anxious or depressed
Which scenario do you think is better?	☐	☐

“Examining aspects of health to expand a measure” survey

Please read each of the scenarios below carefully. You would either live in health scenario A for 10 years and then die or live in health scenario B for 10 years and then die.

Please choose which scenario do you think is better by selecting one of the boxes below.

Block 1

Q 1	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed
Which scenario do you think is better?	☐	☐

Q 2	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems sleeping	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Moderate problems sleeping
Which scenario do you think is better?	☐	☐

Q 3	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with my energy levels	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Extreme problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 4	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems hearing	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Unable to hear
Which scenario do you think is better?	☐	☐

Q 5	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with your social relationships	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - No problems with your social relationships
Which scenario do you think is better?	☐	☐

Q 6	*Health scenario A*	*Health scenario B*
	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with your social relationships	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Moderate problems with your social relationships
Which scenario do you think is better?	☐	☐

Q 7	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with remembering things	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Moderate problems with remembering things
Which scenario do you think is better?	☐	☐

Q 8	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with my energy levels	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Extreme problems with my energy levels
Which scenario do you think is better?	☐	☐

Block 2

Q 1	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed
Which scenario do you think is better?	☐	☐

Q 2	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems hearing	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Moderate problems hearing
Which scenario do you think is better?	☐	☐

Q 3	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems sleeping	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - No problems sleeping
Which scenario do you think is better?	☐	☐

Q 4	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with your social relationships	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Unable to have social relationships
Which scenario do you think is better?	☐	☐

Q 5	*Health scenario A*	*Health scenario B*
	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems hearing	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Moderate problems hearing
Which scenario do you think is better?	☐	☐

Q 6	*Health scenario A*	*Health scenario B*
	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with my energy levels	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Moderate problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 7	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with remembering things	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Moderate problems with remembering things
Which scenario do you think is better?	☐	☐

Q 8	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with remembering things	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Unable to remember things
Which scenario do you think is better?	☐	☐

Block 3

Q 1			*Health scenar*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed	
Which scenario do you think is better?	☐	☐	

Q 2	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems hearing	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - No problems hearing
Which scenario do you think is better?	☐	☐

Q 3	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems hearing	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems hearing
Which scenario do you think is better?	☐	☐

Q 4	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems sleeping	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - No problems sleeping
Which scenario do you think is better?	☐	☐

Q 5	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems sleeping	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Extreme problems sleeping
Which scenario do you think is better?	☐	☐

Q 6	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with my energy levels	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Moderate problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 7	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with your social relationships	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Moderate problems with your social relationships
Which scenario do you think is better?	☐	☐

Q 8	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with remembering things	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - No problems with remembering things
Which scenario do you think is better?	☐	☐

Block 4

Q 1	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems hearing	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - No problems hearing
Which scenario do you think is better?	☐	☐

Q 2	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems sleeping	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Extreme problems sleeping
Which scenario do you think is better?	☐	☐

Q 3	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with my energy levels	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - No problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 4	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with your social relationships	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with your social relationships
Which scenario do you think is better?	☐	☐

Q 5	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems sleeping	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Extreme problems sleeping
Which scenario do you think is better?	☐	☐

Q 6	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with remembering things	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - No problems with remembering things
Which scenario do you think is better?	☐	☐

Q 7	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with my energy levels	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 8	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with remembering things	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Unable to remember things
Which scenario do you think is better?	☐	☐

Block 5

Q 1	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems hearing	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Unable to hear
Which scenario do you think is better?	☐	☐

Q 2	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems sleeping	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Moderate problems sleeping
Which scenario do you think is better?	☐	☐

Q 3	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with your social relationships	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Moderate problems with your social relationships
Which scenario do you think is better?	☐	☐

Q 4	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems hearing	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Moderate problems hearing
Which scenario do you think is better?	☐	☐

Q 5	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with my energy levels	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - No problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 6	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with remembering things	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with remembering things
Which scenario do you think is better?	☐	☐

Q 7	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with your social relationships	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - No problems with your social relationships
Which scenario do you think is better?	☐	☐

Q 8	*Health scenario A*	*Health scenario B*
	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with my energy levels	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Extreme problems with my energy levels
Which scenario do you think is better?	☐	☐

Block 6

Q 1	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems hearing	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Unable to hear
Which scenario do you think is better?	☐	☐

Q 2	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems sleeping	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems sleeping
Which scenario do you think is better?	☐	☐

Q 3	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with your social relationships	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Unable to have social relationships
Which scenario do you think is better?	☐	☐

Q 4	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with your social relationships	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Unable to have social relationships
Which scenario do you think is better?	☐	☐

Q 5	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems sleeping	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Moderate problems sleeping
Which scenario do you think is better?	☐	☐

Q 6	*Health scenario A*	*Health scenario B*
	- Slight problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - No pain or discomfort - Not anxious or depressed - No problems with remembering things	- No problems in walking about - No problems in washing or dressing yourself - No problems doing your usual activities - Slight pain or discomfort - Slightly anxious or depressed - Unable to remember things
Which scenario do you think is better?	☐	☐

Q 7	*Health scenario A*	*Health scenario B*
	- Moderate problems in walking about - Severe problems in washing or dressing yourself - Moderate problems doing your usual activities - Moderate pain or discomfort - Moderately anxious or depressed - No problems with my energy levels	- Moderate problems in walking about - Moderate problems in washing or dressing yourself - No problems doing your usual activities - Severe pain or discomfort - Slightly anxious or depressed - Moderate problems with my energy levels
Which scenario do you think is better?	☐	☐

Q 8	*Health scenario A*	*Health scenario B*
-	- No problems in walking about - No problems in washing or dressing yourself - Moderate problems doing your usual activities - Slight pain or discomfort - Extremely anxious or depressed - No problems with remembering things	- Unable to walk about - Slight problems in washing or dressing yourself - Slight problems doing your usual activities - No pain or discomfort - Not anxious or depressed - Moderate problems with remembering things
Which scenario do you think is better?	☐	☐