



The
University
Of
Sheffield.

An investigation into the psychometric performance of
existing measures of health, quality of life and wellbeing
in older adults

By:

Hannah Penton

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

Date of submission: January 2019

The University of Sheffield
Faculty of Medicine, Dentistry and Health
School of Health and Related Research

Table of Contents

List of abbreviations	viii
Declaration	ix
Acknowledgements	x
Publications and conference presentations	xi
Publications	xi
Conference presentations	xi
Abstract	xii
Chapter 1	1
Introduction.....	1
1.1 Introduction	1
1.2 Aims and objectives	4
1.3 Structure of the thesis	5
Chapter 2	7
Background	7
2.1 Introduction	7
2.2 Economic evaluation in healthcare	7
2.2.1 QALYs	9
2.3 What is the “Q” in QALY?	12
2.4 Arguments for broadening the QALY beyond health	21
2.4 How do we measure QoL? - PROM development	23
2.5 Validation of PROMs.....	27
2.5.1 Psychometric measurement properties.....	27
2.5.2 Psychometric methods for assessing measurement properties.....	30
2.6 Conclusion	46
Chapter 3	47
Selection of existing PROMs potentially suitable for evaluating health and social care interventions and investigation of their psychometric performance in older adults	47
3.1 Introduction	47
3.2 Identification of relevant PROMs.....	47
3.2.1 Methods for identifying relevant PROMs	47
3.2.2 NICE and policy documents.....	48
3.2.3 Rapid review of PROMs which have been used in evaluations of integrated health and social care interventions	50
3.2.4 Researcher experience.....	58

3.3 Selection of PROMs for inclusion.....	59
3.4 Development of included PROMs	63
3.4.1 EQ-5D	63
3.4.2 ASCOT.....	65
3.4.3 WEMWBS / SWEMWBS.....	68
3.4.4 ONS-4	70
3.4.5 SF-12v2	72
3.5 – Systematic review investigating the psychometric performance of the EQ-5D, SF-12, ASCOT, WEMWBS and ONS-4 in older adults.....	75
3.5.1 Introduction.....	75
3.5.2 Systematic review question	75
3.5.3 Methods	75
3.5.4 Results	81
3.5.5 Discussion.....	102
3.5.6 Conclusion	107
3.6 Rationale and research gap	107
3.7 Study aims and objectives.....	110
3.8 Study design.....	111
3.9 Conclusion	113
Chapter 4	114
An investigation into the psychometric performance of the EQ-5D-5L, SF-12v2, ASCOT, ONS-4 and WEMWBS in older adults using item response theory methods	114
4.1 Introduction	114
4.2 Aim	114
4.3 Methods.....	115
4.3.1 Choice of psychometric theory	115
4.3.2 Choice of age cut-off	116
4.3.3 Data sources.....	116
4.3.4 IRT analyses	118
4.3.5 DiF analyses	122
4.3.6 Data preparation	128
4.4 Results.....	129
4.4.1 Sample characteristics.....	129
4.4.2 Response distributions and missing data.....	131
4.4.3 IRT model comparisons and assumption checking	137
4.4.4 DiF model parameters and DiF impact.....	145

4.4.5 Validation	177
4.4.6 Sensitivity analyses.....	184
4.5 Discussion.....	187
4.5.1 Strengths/key findings	187
4.5.2 Limitations.....	192
4.6 Conclusion	193
4.7 Glossary of terms.....	194
Chapter 5	196
Qualitative investigation into the content validity of currently used health and wellbeing measures in older adults	196
5.1 Introduction	196
5.2 Aims and objectives	197
5.3 Methods.....	198
5.3.1 Study design - Choice of data collection method	198
5.3.2 Selection of measures	207
5.3.3 Recruitment strategy and sample size	208
5.3.4 Analysis method.....	211
5.3.5 Reflexivity.....	214
5.4 Results.....	215
5.4.1 Recruitment and respondent characteristics.....	215
5.4.2 Findings	219
5.5 Discussion.....	273
5.5.1 Summary of key findings.....	273
5.5.2 Limitations.....	281
5.6 Conclusion	283
Chapter 6	285
Discussion	285
6.1 Introduction	285
6.2 Integrated key findings.....	286
6.3 Contributions to existing knowledge	294
6.3.1 Systematic review	294
6.3.2 Investigation of psychometric performance using item response theory methods	296
6.3.3 Qualitative investigation of content validity.....	299
6.3.4 Thesis as a whole.....	300
6.4 Implications of thesis findings	301

6.4.1 Implications and recommendations for economic evaluation of health and social care interventions	302
6.4.2 Implications and recommendations for measure development	307
6.5 Limitations	309
6.6 Future research.....	311
6.7 Conclusion	315
References	318
Appendices.....	337
Chapter 3 appendices.....	337
Appendix 1 – Search strategy for rapid review of PROMs in integrated health and social care evaluations in MEDLINE	337
Appendix 2 – Included integration schemes and generic PROMs used.....	338
Appendix 3 – Search strategy systematic review of psychometric evidence of included PROMs in MEDLINE	339
Appendix 4 – COSMIN checklist	340
Appendix 5 – List of included studies.....	351
Appendix 6 – Characteristics of included studies	354
Appendix 7 – Methodological quality of included studies.....	359
Appendix 8– Included study results	361
Chapter 4 appendices.....	377
Appendix 9 – Original response distributions of ONS-4.....	377
Appendix 10 – ScHARR Ethics approval letter.....	378
Appendix 11 – HIPO development and validation sample characteristics	379
Appendix 12 – HSE development and validation sample characteristics	381
Appendix 13 – ASCS development and validation sample characteristics	382
Appendix 14 – EQ-5D IRT MPlus model input files for each stage of analysis plus final model for other measures	383
Appendix 15 – Model fit statistics for each model tested in the DiF identification process	384
Appendix 16 – ICCs for all items.....	391
Appendix 17 – Expected item and measure scores by age group	398
Appendix 18 - Expected item score figures for under and over 65s	407
Appendix 19 – IRT parameters from final DIF model from development and validation samples.....	413
Appendix 20 - Expected item scores from development and validation samples	419
Appendix 21 – IRT sensitivity analysis - Over 75 model parameters	425
Chapter 5 appendices.....	429
Appendix 22 – Demographic questionnaire.....	429

Appendix 23 - NHS Research Ethics Committee and Health Research Authority Approval Letters.....	430
Appendix 24 – University sponsorship letter	442
Appendix 25 – Participant information sheet.....	445
Appendix 26 – Consent form	449
Appendix 27 – EQ-5D concept guide.....	450
Appendix 28 – WEMWBS partial concept guide	452

List of abbreviations

- ASCOT – Adult Social Care Outcomes Toolkit
- BCF – Better Care Fund
- CBA – Cost Benefit Analysis
- CEA – Cost-Effectiveness Analysis
- CFA – Confirmatory Factor Analysis
- CFI – Comparative Fit Index
- CUA – Cost Utility Analysis
- CTT – Classical Test Theory
- DiF – Differential Item Functioning
- EFA – Exploratory Factor Analysis
- GRM – Graded Response Model
- HRQoL – Health Related Quality of Life
- ICC – Item Characteristic Curve
- ICECAP - ICEpop CAPability measures
- IRT – Item Response Theory
- MI – Modification Indices
- ML – Maximum likelihood
- NHS – National Health Service
- NICE – National Institute for Health and Care Excellence
- ONS – Office of National Statistics
- PROM – Patient Reported Outcome Measure
- QALY – Quality Adjusted Life Year
- QoL – Quality of Life
- RMSEA – Root Mean Square Error of Approximation
- SCRQoL – Social Care Related Quality of Life
- SD – Standard Deviation
- TTO – Time Trade Off
- VAS – Visual Analogue Scale
- WEMWBS – Warwick Edinburgh Mental Wellbeing Scale (SWEMWBS – Short WEMWBS)
- WLSMV – Weighted Least Squares Mean and Variance

Declaration

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

I held primary responsibility for all aspects of the research in this thesis. My supervisors, Dr Tracey Young, Dr Claire Hulme and Christopher Dayson, provided guidance throughout the PhD process. They contributed to the design and interpretation of the studies reported in Chapters 3, 4 and 5, and should be considered secondary co-authors of these chapters. Dr Steven Ariss also acted as a mentor for the qualitative study reported in Chapter 5 and provided guidance on its design and interpretation.

Acknowledgements

This project was jointly funded by the NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Yorkshire and Humber, the University of Sheffield and the White Rose Fund. Without this funding this project would not have been possible, and I am extremely grateful for the support of these institutions.

This research also would not have been possible without the kindness and generosity of my participants who took the time to welcome me into their homes and allowed me to ask them too many questions. I thoroughly enjoyed the time spent with each of them and will always be grateful for their help.

I am immensely thankful to my supervisors Tracey Young, Claire Hulme and Christopher Dayson for their insightful suggestions and useful comments throughout the PhD. I am also very grateful for their positivity, which often served as vital encouragement when I most need it.

I am also thankful to the other members of staff who helped me at various stages during the PhD, including Tessa Peasgood and Sarah Pearson for their incredibly useful suggestions during my confirmation review and Clara Mukuria and John Brazier who provided useful guidance during the early stages of this project. I am also hugely grateful to Steve Ariss and Anju Keetharuth for going to the huge effort of reading the whole thesis and providing me with a mock viva. I am also thankful to Steve Ariss for acting as a mentor for the qualitative study presented in Chapter 5.

I would also like to thank those who have been close to me during this process. Sophie, Rachel and Beckie, you (and the cake that often accompanied you) made the journey a little lighter. I could not have done it without your support and friendship. The sense of community within the ScHARR PGR students often made what could have been a lonely process, enjoyable and I will always think of our time together fondly.

Huge thanks also go to my family. To Mum and Dad who always made me feel that anything could be achieved and to Ella who always reminded me to enjoy myself along the way.

Last, but never least, I would like to thank Aure for supporting me, and sometimes dragging me, through this process. I cannot imagine what this journey would have been like without your endless encouragement, patience and affection. Whatever is happening you always provide the perfect escape and I am eternally grateful.

Publications and conference presentations

Publications

Penton H, Young T, Dayson C, Hulme C. (2018) An investigation into the validity of the EQ-5D-5L, SF-12, ASCOT and WEMWBS in older people using item response theory and differential item functioning. *Quality of Life Research*. 27: S62

Penton H, Young T, Dayson C, Hulme C. (2018) Content validity of the EQ-5D-5L, SF-12, WEMWBS and ONS-4 in older people. *Value in Health*. 21: S221

Conference presentations

Penton H, Young T, Dayson C, Hulme C. Content validity of the EQ-5D-5L, SF-12, WEMWBS and ONS-4 in older people. Poster presentation at the ISPOR 21st Annual European Congress, Barcelona, Spain, November 2018

Penton H, Young T, Dayson C, Hulme C. An investigation into the validity of the EQ-5D-5L, SF-12, ASCOT and WEMWBS in older people using item response theory and differential item functioning. Poster presentation at the ISOQOL 25th Annual Conference, Dublin, Ireland, October 2018.

Penton H, Young T, Dayson C, Hulme C. Considerations in conducting economic evaluations in health and social care services in elderly populations: An investigation into the validity of the EQ-5D-5L, SF-12, ASCOT and WEMWBS in older people using item response theory and differential item functioning. Oral presentation at the 3rd Advances Patient Reported Outcomes (PROMS) Conference, Birmingham, United Kingdom, June 2018.

Penton H, Young T, Dayson C, Hulme C. Considerations in conducting economic evaluations in health and social care services in elderly populations: An investigation into the validity of the EQ-5D-5L, SF-12, ASCOT and WEMWBS in older people using item response theory and differential item functioning. Paper discussed at the Summer 2018 Health Economists' Study Group (HESG), Bristol, United Kingdom, June 2018.

Penton H, Young T, Dayson C, Hulme C. An investigation into the validity of existing quality of life and wellbeing measures in older people using item response theory and differential item functioning. Oral presentation at the ScHARR PGR Conference, May 2018.

Penton H, Young T, Dayson C, Hulme C. An investigation of the validity and internal reliability of the ONS-4 Personal Wellbeing Questions in older people. Poster presented at the ScHARR PGR Conference, May 2017

Abstract

The UK population is ageing, with the proportion of the population aged over 65 continuing to rise. Old age is associated with increasing prevalence of frailty, characterised by a slow and steady decline in health and functioning. Older adults experiencing frailty often have complex needs, requiring complicated combinations of health and social care services over the long-term. This has led to a push for integration between health and social care services. Traditionally, the economic evaluation of health services has focussed on health-related quality of life (QoL) as the unit of benefit. However, current health measures may not be appropriate to evaluate the outcomes of social care or integrated health and social care interventions. Outcome measurement methods may need to be adapted to include broader assessment of QoL or wellbeing.

This thesis aims to examine the psychometric performance of existing measures of health, QoL and wellbeing in older adults. First, the existing evidence of the psychometric performance of the EQ-5D-5L, SF-12v2, ASCOT, WEMWBS and ONS-4 was systematically reviewed. Then, item response theory (IRT) was used to examine the construct and structural validity and internal consistency of these measures in older adults aged 65+. Differential item functioning (DiF) analyses assessed whether older and younger adults with the same underlying health, QoL or wellbeing had different expected scores, indicating bias due to age. Lastly, the content validity of the EQ-5D-5L, SF-12v2, WEMWBS and ONS-4 was explored using cognitive interviews in older adults.

This thesis identified some key findings which can inform the choice of measure in the evaluation of health and social care interventions aimed at older adults. In the IRT study, the ASCOT and EQ-5D-5L displayed ceiling effects, while the SF-12v2 and EQ-5D-5L exhibited DiF, both of which can bias the estimates of effectiveness obtained from these measures in an economic evaluation. The cognitive interviewing study provided insight into the way older adults conceptualise QoL and how this impacts the way they respond to items. Issues with response shift were broadly identified, which are the likely cause of DiF. Participants found the functional focused EQ-5D-5L items easier to answer and mostly relevant to their situation. The relevance of broader subjective wellbeing items on the WEMWBS and ONS-4 and negatively phrased mental health items across the measures were commonly questioned as these concepts were not prioritised in older adults' conceptualisation of QoL. However, the coverage of any of the measures would need to be extended to include broader elements of QoL identified as important to older adults. This may be through adaptations to the EQ-5D-5L, such as bolt ons, or the development of a new measure, possibly based on the style of the EQ-5D-5L.

Chapter 1

Introduction

1.1 Introduction

The need for economic evaluations of healthcare interventions is borne out of the fact that demands for healthcare far outweigh the limited public resources available to provide it (Brazier, Ratcliffe et al., 2007). Therefore, if we are to attempt to maximise the possible health that can be obtained from a set budget, we must choose those interventions that offer the best value for money in terms of cost-effectiveness (Drummond, Sculpher et al., 2005). In order to make resource allocation decisions at the health system level, we need a single unit of effectiveness that is comparable across all interventions (Brazier, Ratcliffe et al., 2007). At the general level, healthcare interventions can aim to improve patients in two ways; by extending their length of life and by improving their quality of life (QoL), with many interventions impacting both. Current practice in the UK is to use the quality adjusted life year (QALY) as a generic unit of effectiveness (Brazier, Ratcliffe et al., 2007). The QALY combines the impact of services on the length of life and the QoL of patients so that comparable resource allocation decisions can be made at the National Health Service (NHS) level. However, QoL cannot be directly measured (Fayers and Machin, 2016). Therefore, it is measured using patient reported outcome measures (PROMs). Despite the use of the term QoL, measurement of the QoL element of the QALY to date has focussed on health or health related QoL (HRQoL) in the economic evaluation of healthcare interventions (Brazier and Tsuchiya, 2015, Grewal, Lewis et al., 2006, Milte, Walker et al., 2014).

However, the concepts of health, QoL and related constructs such as wellbeing are not clearly defined within the literature (Fayers and Machin, 2016, Karimi and Brazier, 2016). There is little agreement on what should be included in a measure of health, QoL or wellbeing or even what the differences are between these concepts (Fayers and Machin, 2016). This has led to many PROMs being developed, all based on different definitions of these concepts, containing different dimensions and items. Each measure is built on the assumption that it can comprehensively and feasibly measure the chosen concept in the population in which it will be used and can capture

the impact of any health services they may receive. However, this is not necessarily the case and vigorous testing is required to ensure that the performance of a measure is psychometrically sound, in terms of validity, reliability and responsiveness, in those groups and settings in which it will be used (Fayers and Machin, 2016).

The National Institute for Health and Care Excellence (NICE) help to guide resource allocation in England's NHS by conducting economic evaluations to assess the clinical and cost effectiveness of health interventions and providing recommendations on whether or not these should be provided for free to NHS patients. In 2013, NICE assumed the remit for conducting economic evaluations in social care, to help guide resource allocation decisions in this area of the public sector in the hope of making the most of limited social care budgets (National Institute for Health and Care Excellence, 2016). There is therefore natural interest in extending the current methods of economic evaluation in healthcare to social care. However, the aims of social care interventions are very different to those of health services (Bulamu, Kaambwa et al., 2015, Makai, Brouwer et al., 2014, Milte, Walker et al., 2014, Netten, Burge et al., 2012, van Leeuwen, Jansen et al., 2015). While health services aim to at least maintain, if not improve health, this is often unrealistic in social care, whose aim is more to maintain long-term independence and functioning in the face of impairments which may be stable or worsening (Netten, 2011, Netten, Burge et al., 2012). Social care services often have important outcomes beyond solely health (Brazier and Tsuchiya, 2015) such as independence, dignity and social participation (Makai, Brouwer et al., 2014, van Leeuwen, Jansen et al., 2015), which are often missed by health measures. If these important outcomes are not accounted for in the evaluation of social care interventions there is a risk that these services will be undervalued and underfunded (Brazier and Tsuchiya, 2015, Netten, Burge et al., 2012). Therefore, the current practices of outcome measurement may need to be adapted to include broader measures of QoL or wellbeing.

Older adults, often defined as those aged 65+ (Age UK, 2018a, Age UK, 2018b, World Health Organization, 2002), make up a large proportion of the UK population. The UK population is ageing, with the proportion of the population classed as "older adults on the rise (Age UK, 2018b). It is projected that by 2041, nearly one in four people (24.4%) in the UK will be aged 65+ (Office for National Statistics, 2017b) and the number of people aged 85+ is expected to treble in the next 50 years (Office for National Statistics, 2018). Life expectancies are also increasing. On average, life expectancy at birth rose by 1.4 years for men and 1.0 years for women in the UK

between 2006-08 and 2010-12 (Office of National Statistics, 2014). UK life expectancy estimates at the age of 65 are 85.9 for women and 83.5 for men (Office for National Statistics, 2016b). However, healthy life expectancy is substantially lower at this age. At 65, men in England can expect to live on average another 10.3 years in good health while women can expect another 10.9 years in good health, representing 55.1% and 51.3% of their remaining life respectively (Office for National Statistics, 2016a). As can be expected from these figures, older adults also make up a large proportion of the UK's health and social care service users. It is estimated from the General Lifestyle Survey 2011, that 58% of those aged 65-74 and 68% of those aged 75+ in Great Britain live with a long-standing illness and 36% and 47% of the same age groups respectively reported a limiting long-standing illness (Office for National Statistics, 2013). The ageing population and increasing prevalence of long-term conditions have a significant impact on health and social care spending. Of 18.7 million adult admissions to English hospitals in 2014-15, around 41% were aged 65+ (Hospital Episode Statistics Analysis and Health and Social Care Information Centre, 2015), with length of stay for emergency admissions known to increase with age (NHS Benchmarking Network, 2016). In 2015/16, 51% of social care spending on long and short term support in England was spent on those aged 65 and over (Adult Social Care Statistics and NHS Digital, 2016).

Old age is associated with increasing prevalence of frailty (Gale, Cooper et al., 2015). Frailty is characterised by a slow and steady decline in physical, mental and social functioning, reduced ability to recover from health problems and increased risk of sudden deterioration in health (Clegg, Young et al., 2013, Nicholson, Morrow et al., 2017, Turner, Clegg et al., 2014). Older adults experiencing frailty often have complex needs, requiring complicated combinations health and social care services over the long-term. Those older adults experiencing frailty are at greater risk of disability, care home admission, hospitalisation and death (Fried, Tangen et al., 2001, Rockwood, Mitnitski et al., 2006, Turner, Clegg et al., 2014). The prevalence of frailty is known to increase with age, affecting around 10% of those aged 65+ and increasing to around 65% of those aged 90+ (Clegg, Young et al., 2013, Gale, Cooper et al., 2015).

With older adults representing such an important and increasingly large group in health and social care spending, correctly estimating the most cost-effective services for this group is a priority for the economic efficiency and sustainability of health and social care services in the future. The psychometric performance of those PROMs being used to evaluate services is of utmost importance in this group, however they

are often overlooked in measure development and psychometric testing (Milte, Walker et al., 2014, Ratcliffe, Lancsar et al., 2017). There is also evidence that older populations have different priorities over what is important and relevant to their health and QoL than younger adults (Fayers and Machin, 2016, Ratcliffe, Lancsar et al., 2017). Therefore, this thesis study looks to further explore the psychometric performance of existing health, QoL and wellbeing measures in older adults, in order to examine whether they are suitable for use in the economic evaluation of health and social care services aimed at older adults.

1.2 Aims and objectives

The main aim of this thesis is to assess the psychometric performance of a selection of existing health, QoL and wellbeing measures in older adults, in order to explore whether they are suitable for use in the economic evaluation of health and social care services aimed at older adults. The measures selected are outlined in the objectives below. The methods adopted to select the measures included in this study and the justification of choices made are outlined in Chapter 3. To meet the overall aim of this thesis, the following objectives will be addressed:

1. To systematically review the existing evidence on the psychometric performance of the EQ-5D (3L and 5L) (Brooks and The EuroQol, 1996, Herdman, Gudex et al., 2011), SF-12v2 (Ware, Snow et al., 1993), the Office of National Statistics Personal Wellbeing Questions (ONS-4) (Hicks, Tinkler et al., 2013), the Adult Social Care outcomes Toolkit (ASCOT) (Netten, Burge et al., 2012) and the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) (Tennant, Hiller et al., 2007) in assessing the health, QoL and wellbeing of older adults
2. To examine the structural and construct validity, internal reliability and acceptability of the EQ-5D-5L, SF-12v2, ONS-4, ASCOT and WEMWBS in assessing the health, QoL and wellbeing of older adults using item response theory methods
3. To examine the content validity of the EQ-5D-5L, SF-12v2, ONS-4 and WEMWBS in assessing the health, QoL and wellbeing of older adults using cognitive interviews

4. To provide information on which of the measures tested are appropriate to use in the evaluation of health and social care services aimed at older adults.

1.3 Structure of the thesis

This thesis consists of six chapters. Chapter 2 sets out the relevant background to this thesis. It begins by outlining the process of economic evaluation as the currently used method for allocating limited healthcare resources in the UK. It describes the methods adopted by NICE in the UK, in particular the requirement that cost utility analysis (CUA) economic evaluation be conducted, with outcomes measured in QALYs. The chapter then goes on to examine how QoL is currently measured for use in the QALY in the healthcare sector, and how these methods are being extended given the recent policy shifts in the UK, of NICE extending the methods of economic evaluation to the social care sector and the recent pushes for more integrated health and social care services. The chapter then delves deeper into patient reported outcome measures (PROMs) as the tool for measuring the impact of interventions on the QoL of service users. Definitions and conceptualisations of QoL and related concepts of health and wellbeing seen in the literature are explored and the way that they are being used in practice in economic evaluation is examined. Then the importance of ensuring that the PROMs used to assess the impact of services and make resulting resource allocation decisions are psychometrically sound in terms of whether they are valid, reliable and responsive in the populations in which they are being used is considered. Lastly, the methods that are used to examine the psychometric performance of PROMs are outlined.

Chapter 3 outlines the methods used to identify and select PROMs for inclusion in this thesis study. Choices of measure inclusion and exclusion are justified, before the process by which the selected measures were developed and initially validated is outlined. Chapter 3 then addresses objective one, by presenting a systematic review investigating the existing evidence on the psychometric performance of the chosen measures in assessing the health, QoL and wellbeing of older adults. The systematic methods used to identify relevant existing literature are described, as well as the methods used to assess the quality of identified studies, extract relevant study characteristics and findings and synthesise this evidence across studies. The results

are then outlined and discussed in terms of their strengths and weaknesses. The rationale for the study is then described in the context of the research gap identified in the systematic review. Finally, the aims and objectives for the study, and the design chosen to meet those aims and objectives, are outlined.

Chapter 4 addresses objective two by using item response theory (IRT) methods to examine the structural and construct validity, internal reliability and acceptability of the EQ-5D-5L, SF-12v2, ONS-4, ASCOT and WEMWBS in assessing the health, QoL and wellbeing of older adults. First, the choice to use IRT methods over other available sets of psychometric methods are justified. Then the data sources and methods used are described before the results are presented. These results are then discussed in terms of the strengths and limitations of the analysis undertaken.

Chapter 5 addresses objective three by examining the content validity of the EQ-5D-5L, SF-12v2, ONS-4 and WEMWBS in assessing the QoL and wellbeing of older adults using cognitive interviews. First the choice of data collection methods and selection of included measures are justified and the results of patient and public involvement in the study design are discussed. Then the recruitment strategy, interview protocol and analysis methods are outlined. The results are presented and then discussed in terms of their strengths and limitations.

Finally, in Chapter 6, this thesis is summarised in terms of its aims and objectives. The findings from Chapters 3, 4 and 5 are presented and discussed in terms of how they complement and contrast one another and what evidence they provide in relation to the aims and objectives of the thesis. The contribution of each study to existing knowledge is outlined and the implications of the findings for the measurement of outcomes in older adults and the economic evaluation of health and social care interventions aimed at this population are discussed. Finally, the limitations of the thesis are outlined and recommendations for future areas of research are suggested.

Chapter 2

Background

2.1 Introduction

The aim of this chapter is to provide a detailed background to the topic in order to clearly state why this research is important and necessary. It introduces the current methods used in the economic evaluation of healthcare interventions, with particular focus on CUA and the measurement of HRQoL using the QALY. It introduces the current policy interest in improving integration between health and social care and the impact this has on the current methods of outcome measurement used in economic evaluation, which may need to be extended beyond the measurement of health and HRQoL to broader concepts of QoL and wellbeing. These concepts, their definitions and how they have been conceptualised in the past and present, in relation to the field of health, are then described. The chapter then outlines the process of PROM development, the elements of psychometric performance which should be checked to ensure that a PROM performs well as a measure of the chosen concept in the chosen population and the methods which can be used to examine those elements of psychometric performance.

2.2 Economic evaluation in healthcare

The need for economic evaluation in healthcare is borne out of the fact that healthcare resources are finite. In a given health system there is a set health budget which is only able to provide a certain level of resources. The decision to provide a treatment or service to one group of people with a condition means there are now less resources available to provide for other treatments in other groups (Brazier, Ratcliffe et al., 2007). Therefore, each healthcare resource allocation is associated with an opportunity cost of alternative actions that have to be foregone to be able to provide the chosen service. Therefore, such decisions have an impact on the amount of health that can be generated from that set budget. The process of economic evaluation guides resource allocation decisions in healthcare by assisting decision makers in

making efficient and equitable decisions about the use of resources through the comparison of the costs and benefits of alternative interventions with the aim of maximising the health of the population (Brazier, Ratcliffe et al., 2007).

While it may be fairly straightforward to see how the impact of interventions on the costs of the health system can be assessed, it is much harder to consider how the impact of interventions on health can be assessed (Brazier, Ratcliffe et al., 2007). Different methods of economic evaluation deal with the measurement and valuation of health benefits in different ways (Drummond, Sculpher et al., 2005). In cost benefit analysis (CBA), all benefits, including the impact on health, are valued in monetary terms and then compared to the cost of the intervention. In cost-effectiveness analysis (CEA), health benefits are measured in natural units that are relevant to the decision being made, for example life years saved or increases in bone mineral density (Brazier, Ratcliffe et al., 2007). While this may be adequate when comparing between alternatives that are aiming to achieve the same specific objective, e.g. between different osteoporosis treatments which aim to increase bone mineral density, it cannot be used to compare interventions with different objectives e.g. to compare cancer screening techniques, aiming to increase early detection rates and statins aiming to reduce cholesterol. CEA also cannot be used to compare interventions with more than one outcome, as we cannot obtain a single cost per unit of benefit estimate (Brazier, Ratcliffe et al., 2007). In CUA the health benefits of an intervention are measured in a generic health unit, QALYs, which combine the impact of a treatment on the length of life and the QoL of patients. This enables comparisons to be made between treatments with different specific natural health outcomes and treatments aimed at different populations and conditions, enabling resource allocation decisions to be made at the overall healthcare sector level.

In CUA, the cost-effectiveness of a new treatment is summarised using the incremental cost-effectiveness ratio (ICER). The ICER is calculated as the difference in the cost of a new treatment and current standard care divided by the difference in the effectiveness, measured in QALYs, of the new treatment and the current standard care. This ICER is then compared to a threshold of what the funder is willing to pay per QALY for a new intervention, which provides more benefit to patients and the healthcare system than the previous standard of care, but most often comes at a higher cost.

In England, NICE decide which interventions are funded, based on the results of such economic evaluations. They adopt a threshold of £20,000-30,000 for new treatments and have strict guidelines on the methods to be used in the economic evaluation of health technologies (National Institute for Health and Care Excellence, 2013). NICE requires economic evaluations of health interventions to take the form of CUAs with outcomes measured in QALYs (National Institute for Health and Care Excellence, 2013). In the next sections the QALY and how it is measured is described in more detail.

2.2.1 QALYs

QALYs combine the impact of a treatment on length of life with a utility value for QoL. To date the measurement of QoL for input into the QALY has focused strongly HRQoL (Grewal, Lewis et al., 2006, Milte, Walker et al., 2014). HRQoL is typically measured using standardised PROMs of perceived health status (Drummond, Sculpher et al., 2005). PROMs can either be condition specific, containing domains and items which describe a range of health states relevant to a certain condition or they can be generic, containing more general domains and items relevant across a much broader range of conditions or health issues which an individual might experience. PROMs can be preference-based, meaning they are capable of providing a utility value for the health states described within the PROM, or non-preference-based, meaning they are simply capable of producing a measure of health status but no valuation of that status. In order to be used in the QALY model PROMs must be preference-based (Brazier, Ratcliffe et al., 2007). These preference-based measures (PBMs) include both a descriptive system of potential health states, which forms a questionnaire covering domains, or aspects, of health considered important for the specific condition or health in general and a valuation set for the corresponding health states, valued by the preferences of the general population. This values each health state on a utility scale where 1 is equivalent to full health and 0 to dead. States worse than dead are given a negative utility value. Utility scores are multiplied by the length of time spent in that state in order to produce QALYs.

These preference-based valuation sets are obtained from the general population using valuation methods such as visual analogue scales (VAS), standard gamble (SG), time trade-off (TTO) and discrete choice experiments (DCE) (Drummond, Sculpher et al., 2005). VAS methods involve providing respondents with hypothetical

health states based on the descriptive system of the measure and asking them to first rank them and then to place those hypothetical states, as well as death, relative to each other on a scale between 0, which corresponds to death, and 100, which corresponds to perfect health. SG methods involve offering the respondent a choice between two alternatives (Drummond, Sculpher et al., 2005). Alternative one is a treatment with two potential outcomes and corresponding probabilities; the patient will either be returned to full health and live for another t years, with probability p of occurrence, or the patient dies immediately, with probability $(1-p)$. Alternative two is living for t years in health state i with certainty. Probability p is varied until the respondent is indifferent between the two alternatives giving a resulting preference score for health state $i=p$ (Drummond, Sculpher et al., 2005). SG usually requires face-to-face interview and visual aids, as probabilities are often hard for respondents to relate to.

The TTO process also involves asking respondents to choose between two alternatives. Either to live in state i for time t , followed by immediate death, or to live in full health for time $x < t$, followed by immediate death. Time x is varied until respondents are indifferent between the two alternatives and the preference score for state $i=x/t$ (Torrance, Thomas et al., 1972). Again, TTO exercises are usually conducted using face-to-face interviews, as they can be difficult for respondents to understand. However, more recently online methods have been explored.

Finally, the DCE process involves constructing a series potential health states using the descriptive system of the measure, which capture a broad range of the possible health states which could be described by the measure. Participants are then presented with two or more health states and asked to select which they prefer. The duration of a health state can be incorporated into DCE in order to anchor responses onto the 0=death 1=perfect health scale required for the QALY (Bansback, Brazier et al., 2012). Participants repeat this process multiple times for different combinations of health states. The responses of many participants can then be modelled to reveal their preferences for different domains and levels of items within the PROM. DCEs can be conducted either face-to-face or online.

These valuation methods vary according to whether they account for opportunity costs, the value of the alternative foregone, and individual's risk preferences (Drummond, Sculpher et al., 2005). These concepts are argued to be important for valuation methods to reflect real world choices. VAS, being the quickest and easiest

technique is often criticised as being unrealistic as it fails to account for both opportunity cost and risk preferences. TTO accounts for opportunity cost in asking the respondent to trade years in full health for a set time in state i, however risk attitudes are not examined. Standard gamble accounts for both opportunity cost and risk preferences, as this time individuals are asked to trade off probabilities of successful treatment versus death. An issue with TTO and SG exercises is that they are cognitively demanding and there are concerns that they are not fully understood by less cognitively able groups. DCE exercises as ordinal ranking exercises, are argued to be less cognitively demanding (Bansback, Brazier et al., 2012).

For preference-based PROMs to give accurate estimates of the effectiveness of interventions in a CUA they need to be valid, reliable and responsive (Fayers and Machin, 2016). Validity describes the extent to which the instrument measures what it is aiming to measure. The domains need to comprehensively and appropriately cover the aspects of health which are relevant to the target population and may be affected by any intervention they may receive (content validity), while remaining a feasible length. The relationship between domains should also follow expected patterns according to characteristics such as patient severity (construct validity). A measure also needs to be responsive to health issues which may be experienced in different condition groups or changes which occur in response to treatment, returning appropriately different values which reflect either stability, improvement or worsening of the condition (Brazier, Rowen et al., 2012). PROMs also need to be reliable, meaning that a respondent will return the same utility value where no change in health has occurred.

There are a range different PBMs of HRQoL. These have different descriptive systems, covering different domains, described by different items and have been valued using different valuation methods. These differences mean that different instruments provide different values for the same patient in the same health state (Brazier, Ratcliffe et al., 2007). When allocating resources across the health sector, NICE want to know which interventions are most cost-effective. If effectiveness varies by not only the quality of interventions but also by the PROM used to measure their effectiveness, then this cannot be done reliably. In order for the effectiveness of different interventions to be fairly compared, NICE recommend that all economic evaluations in healthcare gather data on HRQoL using the same instrument. This requires that the chosen measure be generic, capable of validly, reliably and

responsively assessing the impact of the full range of healthcare interventions across populations of patients with different conditions.

The chosen instrument is the EQ-5D-3L (Brooks and The EuroQol, 1996), a generic PBM of HRQoL. It measures HRQoL across 5 domains; mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each domain contains one item with three possible response levels; “no problems”, “some problems” and “extreme problems”/ “unable to”. The EQ-5D also contains a visual analogue scale (EQ-VAS) which provides a self-reported quantitative measure of general health. The EQ-5D-3L has an accompanying value set, valued using TTO exercises in members of the UK general population (Dolan, 1997). More recently the EQ-5D-5L has been developed, covering the same domains but with each item having five possible response levels; “no problems”, “slight”, “moderate”, “severe” and “extreme” problems/“unable to” complete an activity (Herdman, Gudex et al., 2011). An interim mapping algorithm was developed in order to generate utility values for the EQ-5D-5L descriptive system based on the EQ-5D-3L value set, until a EQ-5D-5L value set could be generated (van Hout, Janssen et al., 2012). A UK value set was published in 2017, developed using a combination of TTO and DCE methods in members of the English general population (Devlin, Shah et al., 2017). As long as the EQ-5D is appropriate, in terms of validity, reliability and responsiveness, in the conditions being evaluated, it should enable reliable and comparable benefit calculations in resource allocation decisions across different areas of the health sector. If the EQ-5D is found to be inappropriate, in terms of responsiveness or content validity where it misses key health domains in the relevant population, and evidence is provided, NICE will accept use of alternative PROMs (National Institute for Health and Care Excellence, 2013). There have been conditions and populations where the validity of the EQ-5D has been questioned (Finch, Brazier et al., 2018).

2.3 What is the “Q” in QALY?

As detailed in the previous section the “Q” in the QALY is supposed to incorporate a measure of QoL into the assessment of the benefit of health interventions. This is to allow for the fact that some treatments may lengthen the life of patients, but not necessarily improve their QoL, while other treatments may improve the QoL of patients but may not necessarily lengthen their life (Karimi and Brazier, 2016). However, to date the measurement of QoL in the QALY has focussed on HRQoL.

QoL and related terms such as health and HRQoL are poorly defined and often used interchangeably in the area of health economics causing some confusion (Karimi and Brazier, 2016). A variety of definitions exist for each concept and there is little agreement on what each means, what should be included and how they relate to each other or can be distinguished from one another. This is explored in this section.

There have been many attempts to define and conceptualise health. It is widely acknowledged to be a complex concept which is subject to change (Larson, 1999). Over time, four main conceptual models of health have emerged, summarised by Larson (Larson, 1999). These are the medical model, the wellness model, the environmental model and the WHO model. Traditionally, the most widely used model of health was the medical model (Larson, 1999). The medical model is strongly focussed on defining health as the absence of disease, disability or infirmity (Larson, 1999, Wood, 1986). Wood argues that while health is a relative concept that is virtually impossible to define; illness is an individual's perception that they are suffering from a disease and the consequences of that disease, which can be measured (Wood, 1986). This model focusses on treating specific biological problems as they emerge and has been criticised for being difficult to adapt to mental disorders and for failing to account for preventative medicine and causes of health issues beyond biology, such as social and economic factors, which are known to affect health (Culyer, 1983, Larson, 1999). However, supporters of the medical model view the focus on objective measures of illness as superior to attempts to conceptualise and measure more subjective wellbeing concepts.

The environmental model conceptualises health as an individual's adaptation to the physical and social environment and their ability to perform the roles and activities demanded by that environment while maintaining a balance, free from undue pain, discomfort, disability or limitations related to social abilities (Goldsmith, 1972, Larson, 1999, Navarro, 1977, Parsons, 1972). Therefore, when an individual thrives within their environment and related roles, they are healthy, while a mismatch between the demands of the environment and an individual's ability to function signifies ill health (Abanobi, 1986, Verbrugge and Jette, 1994). This model works well with the concept of health promotion interventions, which aim to use policy to create supportive environments, improve education around health practices and reduce health risks in order to enable individuals to thrive easier within their environment (Larson, 1999, Noack, 1987, Speller, Learmonth et al., 1997). Critics of the environmental model argue that the concepts included in the model would be too broad to be able to

measure (Goldsmith, 1972, Larson, 1999). Also the focus on the environment over the individual may lead to situations where an individual in the same state of health is judged to be healthy in one environment and not healthy in another (Larson, 1991).

The wellness model defines health in terms of health promotion and progress towards higher levels of functioning, energy, comfort, the ability to perform activities and roles and the integration of body, mind and spirit (Dubos, 1979, Greer, 1986, Larson, 1999, Neilson, 1988, Schroeder, 1983). In the wellness model, health is defined as having the strength and ability to overcome illness and recognises the link between the mind and the body (Dubos, 1979, Larson, 1999). It is recognised that individuals have to actively seek recovery and promote health (Dubos, 1979), and emphasises that wellness, recovery and cure are not solely the responsibility of healthcare professionals. Health and illness are also recognised as separate dimensions and not solely opposite concepts (Williams, 1993), in the sense that one can experience bad health, without disease and can experience low burden disease, yet still be healthy. Unlike the medical model, health promotion and prevention are prioritised in this model. However, this model has been criticised for relying on subjective perceptions of health and wellness, which are known to vary by age and cultural context (Larson, 1999). Other critics add that this model expands the concept of health to include happiness and QoL, which means that someone may be healthy according to the medical model, but unhappy and have a low QoL according to the wellness model (Larson, 1991).

The WHO model is currently the most widely used definition of health. The more holistic definition was proposed by the World Health Organization (WHO) in 1948, which stated that health is “a state of complete physical, mental and social well-being, and not merely the absence of disease” (World Health Organization, 2015). This definition encourages the consideration of individuals as “social beings whose health is affected by social behaviour and interaction” (Larson, 1999). When released, it was felt that this concept was both idealistic and unmeasurable (Larson, 1999). However, since its conception substantial research has gone into developing measures based on this holistic definition, such as the SF-36, SF-12 and EQ-5D measures (Ware, Snow et al., 1993). Criticism of the WHO model includes arguments that, while social factors may directly affect health, there is a lack of evidence that social wellbeing is an independent domain of health and therefore social wellbeing should not be used to define an individual’s health status (Ware, Brook et al., 1981). Other criticisms include the difficulty of accurately measuring such broad and ill-defined concepts as

wellbeing within health (Bice, 1976, Pannenborg, 1979). Despite these arguments the WHO model has become the most used definition in the world, providing a very broad definition of health. However, this definition of health may not be vastly distinguishable from how we would define QoL or wellbeing.

Similarly to health, there have been many attempts to conceptualise and define QoL, often based on very different approaches, which have led to a wide range of possible definitions (Felce and Perry, 1995, Ferrans, 1990). More basic definitions and conceptualisations of QoL have focussed on there being an objective list of conditions which, if met, mean we would view an individual's QoL positively (Brazier, Connell et al., 2014). However, this has been criticised as it ignores individuals' preferences and values in terms of what they consider to be a desirable life. In response to this, it has been argued that QoL should be defined as a combination of the quality of an individuals' life conditions, their satisfaction with those life conditions and their personal values, aspirations and expectations (Felce and Perry, 1995). In this definition, the concepts of QoL and life satisfaction are considered separate domains, which are closely interwoven. According to this model, the conceptualisation of QoL should include a series of objectively measured life conditions such as physical health, personal circumstances (wealth, living conditions, etc.), social relationships, functional activities and pursuits, and wider societal and economic influences (Felce and Perry, 1995). Individuals' subjective response to the above aspects of QoL would then indicate how satisfied they are with these conditions. The extent to which individuals feel satisfaction towards different levels of different conditions would be based on their personal values, aspirations and expectations (Felce and Perry, 1995). This led to the following definition of QoL as "an overall general well-being that comprises objective descriptors and subjective evaluations of physical, material, social and emotional well-being together with the extent of personal development and purposeful activity, all weighted by a personal set of values" (Felce and Perry, 1995).

The above definition closely links with other definitions of QoL, which also emphasise the importance of not only objective conditions under which a life may be viewed as good, but also individuals' personal preferences and satisfaction with their situation. For example, Rejeski defined QoL as "a conscious cognitive judgement of satisfaction with one's life" (Rejeski and Mihalko, 2001). It is also commonly acknowledged that what may be viewed objectively as a good life, as well as peoples' perception and satisfaction with their own situation, is likely to be influenced by culture specific values. This was acknowledged in the definition by Kuyken who defined QoL as "an

individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns" (Kuyken and Group, 1995). These definitions of QoL, which are commonly conceptualised with reference to wellbeing and life satisfaction, closely align the concepts of QoL and wellbeing.

Again, there is a lack of agreement on the definition of wellbeing. As we have seen, this term is often included within definitions of health and QoL, without being clearly defined or distinguished. Broadly, wellbeing has been defined as how well an individual's life is going (Brazier and Tsuchiya, 2015, Peasgood, Brazier et al., 2014). However, there are a collection of well-established theories regarding what is important to be able to consider that an individual's life is going well (Peasgood, Brazier et al., 2014), which are often used as the theoretical basis for the measurement of wellbeing. These include objective list theories; preference satisfaction theories; hedonist theories; evaluative theories; and flourishing theories.

Objective list theories, similar to the objective lists used in the conceptualisation of QoL, argue that an individual's life goes well if they have certain goods or characteristics, regardless of whether they want or enjoy them (Peasgood, Brazier et al., 2014). Examples of such goods or characteristics could be access to housing, enough food, being able to see and read, having friends and feeling safe. Again, this is criticised for ignoring what individuals feel is important to their life. By contrast, preference satisfaction or desire theories state that wellbeing increases when an individual gets what they desire (Dolan and Metcalfe, 2012). The issue with this theory is that it is difficult to measure in contexts such as health and social care. In economics, individuals reveal their preferences for goods and services through their willingness to pay for them. However, in the UK health and social care system, such markets do not necessarily exist. Willingness to pay techniques have been used to attempt to estimate individuals' willingness to pay through contingent valuation exercises. However, these techniques are often criticised as individuals' stated willingness to pay for gains in health is strongly associated with their ability to pay which has negative distributional implications (Peasgood, Brazier et al., 2014).

Hedonic theories suggest that an individual's wellbeing increases if they experience more pleasure than pain and therefore depends on an individual's feelings at a given time (Dolan and Metcalfe, 2012, Peasgood, Brazier et al., 2014). Possible questions based on hedonism would ask people about their current positive and negative

feelings such as happiness, enjoyment, worry or sadness. This has been argued to be too simplistic, as positive and negative affect are not the only things that people care about or desire (Peasgood, Brazier et al., 2014). There have also been concerns about individuals' ability to accurately recall past feelings (Kahneman, Wakker et al., 1997, Peasgood, Brazier et al., 2014).

Evaluative theories argue that life goes well if an individual evaluates their life positively. This theory would ask broad questions about life satisfaction, happiness or fulfilment and it would be left up to the respondent to cognitively appraise their life as a whole, choosing to include whatever information they feel is relevant (Peasgood, Brazier et al., 2014). This model, while praised for allowing individuals to evaluate what is important to them, has been criticised due to concerns regarding individuals' ability to accurately evaluate their life using such broad questions (Peasgood, Brazier et al., 2014). Other criticisms include concerns that responses may be overly dependent on the respondent's current mood and context (Peasgood, Brazier et al., 2014, Schwarz and Clore, 2004).

Finally, flourishing theories, or eudemonic theories, suggest that an individual's life goes well if they have the ideal qualities for a human being to have, with wellbeing increasing as they reach their true potential in life and flourish (Peasgood, Brazier et al., 2014). Examples of concepts that would be considered in flourishing theory are autonomy, purpose in life, connectedness and achievement (Dolan and Metcalfe, 2012, Peasgood, Brazier et al., 2014). One possible criticism of this model is that it assumes that people only value the things or activities in their life in relation to their impact on these concepts, regardless of any enjoyment they get from those things or activities (Dolan and Metcalfe, 2012).

Another approach which has more recently been adopted in the definition and conceptualisation of wellbeing has been to define it in terms of individuals' capabilities (Dolan and Metcalfe, 2012, Peasgood, Brazier et al., 2014). The capability approach, grounded in the theories of Sen, argues that what is important to wellbeing is that an individual is able to function in a certain way, regardless of whether they choose to or not (Sen, 1982). For example, it is widely understood that lifestyle choices such as a healthy diet and regular exercise lead to better health outcomes. The capabilities approach would argue that it is not important to an individual's wellbeing whether or not they eat healthily and do regular exercise, but that they are able to do so, if they desire. Sen's theory considers that a person should have the capability to achieve a

basic set of functionings such as “moving, being well nourished, being in good health and being socially respected (Sen, 1982). However, the theory only suggests these as possibly important capabilities and does not provide a firm set of capabilities that would define “good” wellbeing. This has been a criticism of this theory, although proponents of the theory argue that the selection of attributes is required of any theory in this area and is likely to be dependent on the situation (Grewal, Lewis et al., 2006, Robeyns, 2003, Sen, 1983).

The term HRQoL was introduced in relation to the literature on health status measures and QALYs. However, again its definition is unclear and a variety of commonly used definitions exist which have led to the term being used to mean different things within the field of health economics (Karimi and Brazier, 2016). Four such definitions from the literature have been identified and outlined by Karimi and Brazier (Karimi and Brazier, 2016). The first defines HRQoL 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). This definition of HRQoL mirrors the WHO definition of health, as HRQoL is essentially collapsed into functioning and perceived wellbeing in physical, mental and social health. The second definition identified states that “quality of life is an all-inclusive concept incorporating all factors that impact upon an individual’s life. Health-related quality of life includes only those factors that are part of an individual’s health” (Torrance, 1987). This definition argues that HRQoL would exclude non-health aspects of QoL, for example economic and political circumstances. This definition also collapses HRQoL to the concept of health, which is acknowledged to be an element of overall QoL.

The third definition identified, defines HRQoL as “those aspects of self-perceived well-being that are related to or affected by the presence of disease or treatment” (Ebrahim, 1995). This is commonly used to distinguish between those elements of QoL and wellbeing that are of interest in the economic evaluation of health interventions (i.e. those aspects which are impacted by health issues and treatments) and those elements of QoL that are broader, for example life satisfaction and happiness, which may not be judged to be relevant to the economic evaluation of health services (Fayers and Machin, 2016). However, it is not clear where this line should be drawn. While some impacts of illness and treatment may be clear and directly related to health there are many more aspects of life which are indirectly impacted by bad health and subsequent treatment. For example, bad health can lead to an inability to work. This can lead to financial stress and mental health issues and

reduced life satisfaction and happiness. Similarly, poor health can mean people find it hard to get out of the house, leading to social isolation. Therefore, questions about social participation, ability to work or carry out other activities and life satisfaction, which may seem like elements of wider QoL and not directly relevant to health may in fact be relevant to HRQoL. Any distinction also relies on the definitions of health and QoL used, which can vary widely in breadth of coverage also (Brazier, Connell et al., 2014).

The fourth definition represents a distinctly different use of the term HRQoL and is used in the valuation of health to refer to the utility values assigned to health states from preference-based measures (Karimi and Brazier, 2016). Each of these definitions, and the breadth of aspects that they include, again crucially rely on an individual's interpretation of the concepts of health, QoL and wellbeing. Each of the definitions of HRQoL, with the exception of the fourth definition relating the concept to valuation, essentially collapses down to either the definition of health or QoL depending on the definitions of those concepts followed (Karimi and Brazier, 2016).

Despite a lack of agreement on definitions for these terms it is widely accepted that health, QoL, wellbeing and HRQoL are all multidimensional concepts (Fayers and Machin, 2016, Karimi and Brazier, 2016). They may just differ in the breadth or type of dimensions they include. Health is accepted to be an element of QoL and wellbeing (Brazier, Connell et al., 2014, Ferrans, 1990, Makai, Brouwer et al., 2014), with these two broader concepts also capturing elements such as happiness and life satisfaction, which tend not to be included in conceptualisations of health. HRQoL is, in theory, used to select a subset of important or common ways in which health or treatments impact upon QoL or wellbeing. However, the complex relationships between these interwoven concepts means that it is not clear where this line should be drawn. The breadth of the concept of HRQoL is very much dependent on definition of HRQoL chosen as well as the definitions of health and QoL. However, once these choices are made, the definition of HRQoL (except when referring to utility values) seems to essentially collapse into either the definition of health or QoL, making this term a largely irrelevant distinction when used in this way. Therefore, for the purpose of this study and following the recommendations set out in Karimi and Brazier (Karimi and Brazier, 2016), measures which focus on health will be referred to as health status measures, while broader measures will be referred to as QoL or wellbeing measures. The term HRQoL will be used when referring to the utility value associated with health

states, for example when the value set has been used to convert EQ-5D responses into utility values (Karimi and Brazier, 2016).

PROMs are widely recognised as the method for measuring subjective concepts such as health, QoL and wellbeing. The history of PROM development shows a steady trend in recognition of the broad impact that health services have on patients' lives. Traditionally the measurement of health in PROMs took a narrower focus (Fayers and Machin, 2016), with the earliest PROMs focussed on physical functioning and activities of daily living. In the 1970s and 1980s the field evolved to develop generic health status measures, encompassing aspects of health thought to be impacted by health issues and treatments such as physical functioning, psychological symptoms and the impact of illness (Fayers and Machin, 2016). A lack of clear and agreed definition meant there was substantial variety in what aspects of health were included in different measures. The EQ-5D (Brooks and The EuroQol, 1996) and SF-36 (Ware, Snow et al., 1993), two of the most prominent health status measures in use today, were developed in the early 1990s, reflecting the WHO definition of physical, mental and social elements of health. Currently in economic evaluation in the UK, NICE require that the outcomes of health services be measured in QALYs using the EQ-5D measure of HRQoL (National Institute for Health and Care Excellence, 2013) and therefore health status measures represent the current state of play in measurement the impact of treatments in the UK.

However, more recently there has been increasing interest in broadening the QALY beyond solely health towards broader QoL and wellbeing (Brazier and Tsuchiya, 2015). This has been reflected in the generation of many broader instruments, for example the WEMWBS measure of positive mental wellbeing, the ICEpop CAPability (ICECAP) capabilities measures (Al-Janabi, Flynn et al., 2012, Coast, Flynn et al., 2008, Grewal, Lewis et al., 2006), the ASCOT measure of social care related quality of life (SCRQoL) and more recently the Recovering Quality of Life (REQoL) mental health measure (Keetharuth, Brazier et al., 2018). The current policy focus on developing economic evaluation methods in social care and creating more holistic integrated health and social care delivery in individuals with complex needs, such as frail older populations, both create important arguments for broadening the QALY beyond solely health. These arguments will be discussed in the next section.

2.4 Arguments for broadening the QALY beyond health

Several recent policy shifts in the UK have driven the need for broader outcome measures in economic evaluation. The first was that NICE extended its role by assuming the remit for conducting economic evaluation of social care interventions in 2013 (National Institute for Health and Clinical Excellence, 2013). Similar to healthcare funding, there is an obvious desire to get the most benefit for social care service users out of a fixed and very strained social care budget and NICE therefore began making steps to initiate economic evaluation processes for social care interventions. The second important policy shift was the recent UK shift towards integration in health and social care services (Department of Health, 2013, Humphries, 2015). This is in response to the ever-increasing financial pressure on both health and social care services arising from an ageing population. As health technologies improve and patients live longer, they are more likely to experience a greater number of health conditions (Goodwin, Dixon et al., 2014). As the proportion of elderly people within the population increases, so does the number of patients with complex needs (patients with more than one long-term condition) and therefore the number of individuals requiring a mix of health and social care interventions, often over the long-term (Bennett and Humphries, 2014).

Different areas of health and social care services such as acute hospital care, mental health services, community health services and community and residential social care services have become increasingly fragmented with each being organised, commissioned and funded separately, with their own accountability and performance structures (Humphries, 2015). However, for service users such as older adults with complex needs, who require a combination of both health and social care services, the distinction between which body those services are provided by is often unclear (Grewal, Lewis et al., 2006). Lack of understanding of the complex organisational structure of these fragmented services and failures in communication can mean that such patients do not receive optimal and timely care. These factors have resulted in continued calls for a more integrated health and social care system (Grewal, Lewis et al., 2006). The Better Care Fund (BCF), announced in June 2013 (The Better Care Fund, 2014), aimed to encourage and ensure integration by providing local pooled budgets to be spent jointly on partnership working between health and social care services by local authorities and clinical commissioning groups (The Better Care Fund, 2014). However, there was a lack of clear guidance on how BCF projects should be evaluated.

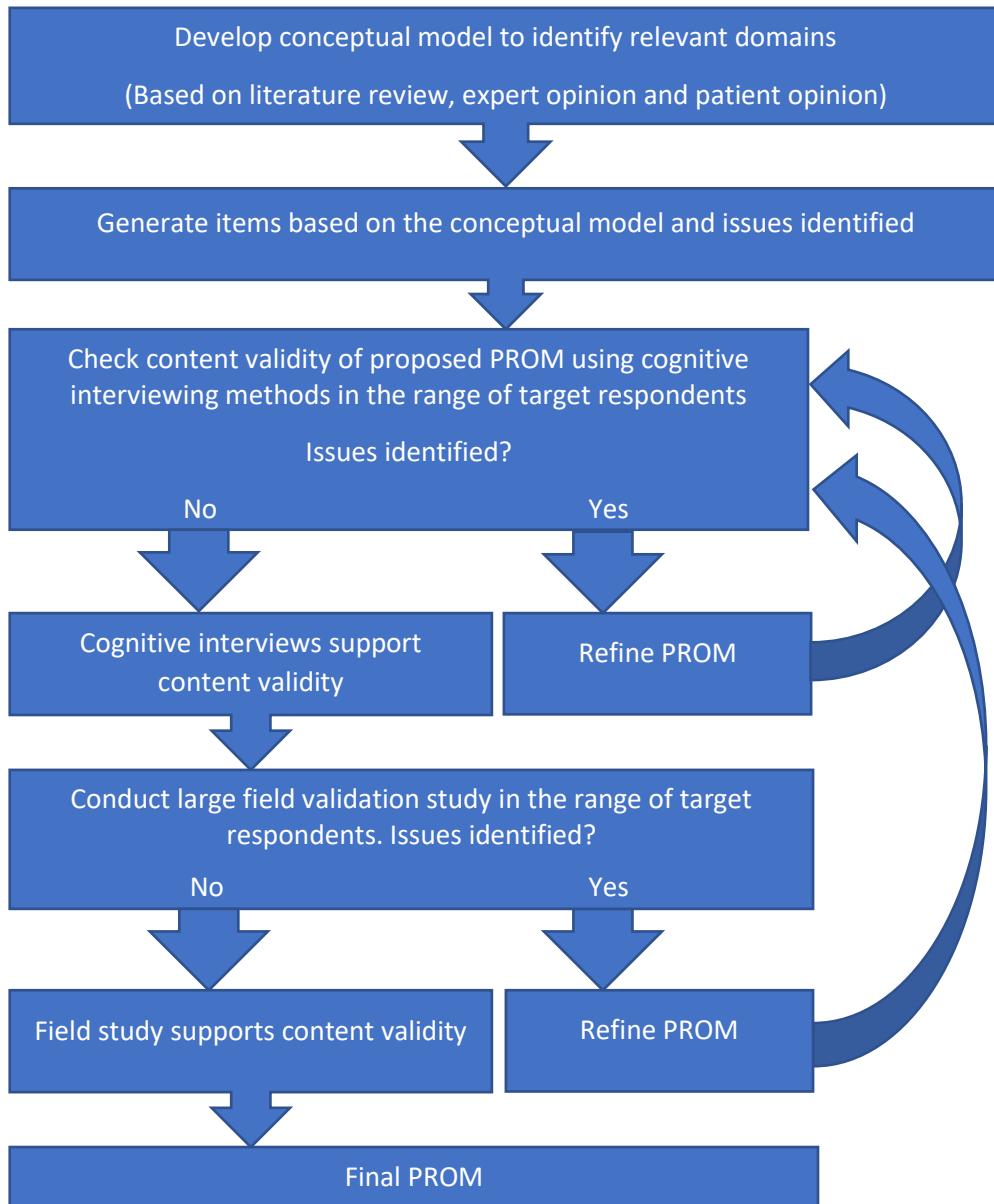
The main aim of the BCF was to reduce the incidence of unplanned admissions to hospital by improving integration and reducing the likelihood that individuals would fall through the gaps between fragmented services. Another of the expected outcomes of integration was a positive impact on QoL (Curry and Ham, 2010). An important issue in the evaluation of integrated services is that the desired outcomes of health and social care services can be quite different (Makai, Brouwer et al., 2014). Health services often aim to improve or maintain health and are often delivered over short periods where needed. However, social care interventions are often long-term services seeking to improve or maintain non-health aspects of QoL such as independence, dignity or comfort (Makai, Brouwer et al., 2014, Netten, Burge et al., 2012, van Leeuwen, Jansen et al., 2015). Therefore, an outcome measure is required which measures aspects of QoL beyond health.

Methods in the economic evaluation of social care interventions are much less developed than in healthcare. Social care methods guidelines have been produced (National Institute for Health and Care Excellence, 2016) but they are much less clear cut than those for healthcare. NICE are explicit that ideally, they would prefer a CUA, which allows effective comparisons across different decision problems (National Institute for Health and Care Excellence, 2016). However, they acknowledge that this may not currently be possible, as methodological issues mean there is currently no accepted social care equivalent to the healthcare QALY. The social care guidance manual states that economic evaluations in social care should measure and value effects using the QALY, or a “social care” QALY with a parallel evaluation based on a capability measure or wellbeing where an intervention has outcomes in health, social care and capabilities. Any preference data for the valuation of changes in QoL should come from a representative sample of the UK public. If health effects are relevant the health QALY, measured using EQ-5D (Brooks and The EuroQol, 1996) is suggested. ASCOT (Netten, Burge et al., 2012) is suggested as a measure of SCRQoL while the ICECAP instruments, ICECAP-A and ICECAP-O (Al-Janabi, Flynn et al., 2012, Coast, Flynn et al., 2008, Flynn, 2011) are suggested as measures of capabilities for adults and older adults respectively. The choice of wellbeing measure is left open. This flexible approach to the measurement of outcomes has issues for the comparability of economic evaluations and resource allocation in social care.

2.4 How do we measure QoL? - PROM development

This section details the stages that must be completed to develop a useful PROM. Figure 1 depicts the stages of the process, which are described in more detail below.

Figure 1 – PROM development process



Define aim of PROM and conceptual model

Before PROM development can begin, important decisions regarding the aim of the PROM, in terms of what it is aiming to measure and who will be the intended respondents must be made (Patrick, Burke et al., 2011a). PROMs can differ greatly on what they aim to measure; whether it is health, QoL or wellbeing. It is essential to

have a clear definition of what it is the PROM is measuring as this will affect what is relevant to include. As previously discussed, there are various possible definitions for each of these concepts and therefore many possibilities for potential conceptual focuses. For example, a health measure could focus on symptoms, physical functioning, mental health or social functioning or a combination of the four (Fayers and Machin, 2016). There is also the important choice of whether the PROM should focus on a specific condition and the issues associated with this, or whether it should be generic, suitable for assessing the impact of a wide range of health issues. In order to make funding decisions across different diseases and populations a generic measure is required to ensure comparability between evaluations.

The characteristics of the intended population, and the conditions and treatments they may have, will have an important impact on the required design and coverage of the measure (Patrick, Burke et al., 2011a). Participants of different ages, cultural backgrounds or experiencing different conditions may have different priorities of what concepts or domains are important to include (Fayers and Machin, 2016). They may also interpret questions differently or have different opinions on what is acceptable to ask. Very young, very old or very ill patients may also need help completing the questionnaire. Questions also need to be relevant and able to discriminate between individuals across the full range of severity of the target population; from very ill to very healthy.

Conceptual Domain and Item Generation

Once the aim of the PROM has been defined, a list of all potentially relevant and important themes to the concept being measured can be generated. This should be done using a combination of searching the existing literature and interviewing relevant experts and members of the target population (Fayers and Machin, 2016, Rothrock, Kaiser et al., 2011). The next stage is to create a list of all issues that could be relevant to the domains of interest. Again, this should draw on the existing literature and interviews with relevant experts and members of the target population. Existing measures which assess the same or related concepts should be reviewed as they are good sources for potentially relevant issues which could be included in the instrument being developed. Once a full list has been generated this should be discussed with relevant experts in this area such as health workers, psychiatrists or social workers. The content validity of the proposed list should be assessed in terms of whether the

issues included are relevant and whether the list comprehensively covers what is important to the concept being measured or whether additional issues need to be added. The revised list should then be discussed with a sample of the target population to further establish content validity in a similar way. This sample should be representative of the full range of respondents expected to receive the PROM in practice (Patrick, Burke et al., 2011a). Respondents with different characteristics (e.g. age or cultural background), different conditions and severities may have different views on what is important to the concept being measured and it is important to capture all these views for the measure to be relevant and comprehensive to the full range of participants (Fayers and Machin, 2016).

Developing Items

The next stage is to convert the resulting list of relevant issues into questions, or items. Many aspects of item design need to be carefully considered, including format, question and response option wording and time frame (Rothrock, Kaiser et al., 2011). Questions can be dichotomous, e.g. yes/no format, or ordinal, with respondents able to mark themselves according to a series of response options on some form of graded scale, e.g. “not at all” to “very much” or “never” to “all of the time”. The number of categories chosen can affect the performance of an item, with less than four suggested to be too few for the item to be able to discriminate well between individuals, while respondents have been shown to struggle to reliably and repeatedly discriminate between categories when there are more than six (Fayers and Machin, 2016).

The wording of questions is crucial to respondents’ understanding and interpretation of them and the resulting validity of their responses. Questions should be brief, clearly worded, easily understood, unambiguous and easy to respond to (Fayers and Machin, 2016). Questions and instructions should be brief and simple as older and severely ill respondents may get easily confused and font should be large and easy to read. Questions should be appropriate and relevant to the range of target respondents. Questions which respondents feel are inappropriate or irrelevant may be left blank by respondents or they may provide inconsistent answers. For example, say a respondent who has mobility problems and is confined to the bungalow is given the question “are you limited in climbing several flights of stairs”. They may feel it is not relevant and leave it blank; they may respond “yes limited” because they would be unable to do it if they tried; or they may answer “no, not limited” because they would

never need to try and therefore do not experience difficulty with this. Questions which appear to meet these criteria and often found to cause unexpected problems in practical use and therefore it is very important that these questions are rigorously tested in members of the target population before they are released for general use (Fayers and Machin, 2016).

A concept, or domain, may be covered by a single question, if the concept is felt to be fairly simple e.g. pain, or by multiple questions if the concept is more complex e.g. depression and it is felt that more questions would achieve a greater precision of measurement of the concept. Items which measure different aspects of a domain are often grouped together to form multi-item scales. It can be useful to score these related items together to form summary scores related to specific domains within the measure.

Face and Content Validity Pre-testing

The proposed PROM should then be tested again by relevant experts and members of the target population in order to ensure that the resulting version has face and content validity and is acceptable and appropriate for the target population. This should include confirming whether the items are relevant and comprehensively cover what is important to the concept being measured and whether the items are clear, easy to understand, unambiguous, appropriate and acceptable. In pre-testing, respondents should be asked to complete the proposed version of the questionnaire and then face and content validity should be assessed using cognitive interviewing methods (Rothrock, Kaiser et al., 2011). More details about these methods will be provided in section 2.5.2.3. Pre-testing should be carried out in a representative sample of those who will receive the PROM in real world practice. Any issues identified with item wording, relevance, appropriateness or insufficient coverage of important concepts should be resolved and pre-tested before a large field test is carried out.

Field Test Questionnaire

Once content and face validity have been confirmed the PROM can now be tested further in a larger group of respondents. The aim of this stage is to use responses to

assess the psychometric performance of the measure and to ensure that it is performing in the way that is intended and required. Aspects of psychometric performance that should be tested and the most commonly used methods for assessing psychometric performance are described in section 2.5. Field testing should be carried out in a large sample representative of the full range of respondents who will receive the measure in practice. This includes the full range of conditions, severities, settings and demographic characteristics, including age and cultural backgrounds as these are all variables that can affect understanding, interpretation, appropriateness and relevance of questions (Fayers and Machin, 2016, Mallinson, 2002). Again, if issues arise that require changes to be made to the measure, the psychometric performance of the revised measure should be retested. If the developers intend for the instrument to be used in different countries the measure will need to be translated, following cross-cultural validation and translation guidelines (Beaton, Bombardier et al., 2000). A debriefing questionnaire should also be included to identify any issues with ease of completion, understanding, acceptability, relevance and comprehensiveness (Fayers and Machin, 2016).

2.5 Validation of PROMs

2.5.1 Psychometric measurement properties

This section will outline the important aspects of the psychometric performance of PROMs, which need to be investigated before we can be sure that a PROM can provide an accurate, precise and unbiased estimate of the underlying trait it is aiming to measure.

2.5.1.1 Validity

Validity is the extent to which an instrument measures what it is intended to measure (Mokkink, Terwee et al., 2010), for example to what extent do a group of items on a PROM, intending to measure pain, actually measure pain. Content validity assesses the degree to which the items adequately reflect the construct being measured (Mokkink, Terwee et al., 2010) with the aim of comprehensively covering the important aspects of the construct while accounting for respondent burden. So, in the pain example, are all the important aspects of pain covered by a feasible number of items.

Face validity, an element of content validity is an assessment of the degree to which the items of a PROM do indeed look as though they are an adequate reflection of the construct being measured (Mokkink, Terwee et al., 2010). Criterion validity is the degree to which the scores of a PROM are an adequate reflection of a “gold standard” (Mokkink, Terwee et al., 2010). So, if there were an objective or known best measure of pain, how well do the scores of this PROM relate to those scores. Construct validity assesses how well the PROM measures what it is intended to measure and checks that the measure performs as expected (Fayers and Machin, 2016) in terms of known group validity, convergent and discriminant validity and measurement invariance. Known group validity examines whether the scores of a PROM can differentiate between groups of respondents who would be expected to score differently (Mokkink, Terwee et al., 2010), for example do those who are known to have different severities of pain receive appropriately different scores on the measure. Convergent and discriminant validity examine whether the scores from the PROM have the expected relationships with scores of other relevant instruments (Mokkink, Terwee et al., 2010), so do scores from this measure of pain have expected relationships with scores from other related measures of pain or health.

Measurement invariance is a key assumption of PROMs, which states that the PROM is measuring the same underlying construct in the same way in all groups of respondents (Putnick and Bornstein, 2016). The relationships between: the items and the latent trait; between the items themselves and between response options should be the same across groups. This is required for valid unbiased group comparisons to be made. Therefore, the items chosen to measure pain, should perform the same in all groups of respondents. The items should have the same relationships with each other and the true underlying pain scale in all groups and all groups should interpret the items and the response options in the same way, so for example “moderate” pain should mean the same level of pain to everyone. Items which perform differently across subgroups exhibit what is called differential item functioning (DiF), which may introduce bias into scoring that will cause issues, especially in relation to group comparisons.

2.5.1.2 Reliability

Reliability is defined as the degree to which a measure is free from error (Mokkink, Terwee et al., 2010). Internal consistency reliability assesses the degree of

interrelatedness among the items of the PROM (Mokkink, Terwee et al., 2010) and therefore the extent to which items on a scale measure the same concept, in our example to what extent all the items on the PROM measure pain. Test-retest reliability examines the extent to which scores in patients who have not changed remain the same over time (Mokkink, Terwee et al., 2010). This is usually measured at two different time points over a reasonable interval of approximately 7-days, where we expect scores to remain stable. Inter-rater reliability measures the stability of a person's score when a PROM is completed by two different raters at the same time (Mokkink, Terwee et al., 2010), so for example if a carer and patient both complete the pain PROM for the patient's level of pain, inter-rater reliability assesses the extent of agreement between the answers they provide.

2.5.1.3 Responsiveness and sensitivity

Responsiveness and sensitivity are two closely related aspects of measurement performance. Responsiveness is the ability of a PROM to detect change over time in the trait being measured if a change exists (Mokkink, Terwee et al., 2010). For example, if the pain measure is completed by a respondent before and after an intervention which successfully reduces their pain, do the scores pick up that improvement. Sensitivity is the ability of the PROM to detect differences between groups, for example between groups with different disease severity or different treatment groups in a trial (Fayers and Machin, 2016). In this case does the pain PROM return different scores for those with different true severities of pain or between different groups in a clinical trial (for example between treatment groups who received an effective pain reducing intervention and a placebo group). A measure which performs poorly on these properties will be unable to detect important change in respondents QoL and unable to discriminate between patients with different levels of QoL, both of which are important.

In the next section the main schools of psychometric methods used to assess these properties will be introduced and the tests and methods used within each school to test the properties will be outlined.

2.5.2 Psychometric methods for assessing measurement properties

In this section, the different schools of psychometric theory are described, and the different tests commonly used within each school, to examine the performance of instruments in relation to each psychometric property are outlined and discussed.

2.5.2.1 Classical test theory

Classical test theory (CTT) is the traditional branch of psychometric testing and remains the most commonly used. CTT is based on the work of Spearman who introduced the decomposition of an observed score from a test into an unobservable true score, which quantifies their level of the underlying trait being measured and is defined as a person's expected score over an infinite number of independent administrations of the measure (Cappelleri, Jason Lundy et al., 2014), and an error (Petrillo, Cano et al., 2015). This led to an interest in the estimation of the reliability of the observed scores (Petrillo, Cano et al., 2015). CTT tests are based on several assumptions. The first assumption is monotonicity, meaning that as true scores increase, so should responses to items representing that concept, assuming that item responses are coded so that higher responses reflect higher levels of the underlying trait (Cappelleri, Jason Lundy et al., 2014). It is also assumed that errors found in observed scores are random and normally distributed with a mean of zero (Cappelleri, Jason Lundy et al., 2014). These random errors are also assumed to be uncorrelated with the true score, and therefore there should be no systematic relationship between a person's true score and error (Cappelleri, Jason Lundy et al., 2014). CTT tests are also based on assumptions that items can be summed to form a total score, without needing to be weighted or standardised (Petrillo, Cano et al., 2015). The tests of validity and reliability are mostly based on descriptive statistics and correlations, as outlined below.

CTT validity tests

CTT tests of construct validity mostly examine the degree to which scores of a PROM are consistent with hypotheses that would be expected from an instrument which validly measures the underlying construct which it is said to measure (Mokkink, Terwee et al., 2010). These hypotheses can concern external relationships between the PROM and other measures or relationships between scores in different groups of respondents. Convergent validity investigates the extent of agreement with other

measures intended to measure the same construct (Haywood, Garratt et al., 2005). In tests of convergent validity, the direction and magnitude of correlations between the scores of different measures are hypothesised a priori and tested to see if the expected relationships hold. A priori hypotheses are also made regarding the expected direction and magnitude of differences in scores between groups who are known to differ on the trait being measured (known group validity) and statistical tests are carried out to examine whether there is a statistically significant difference in the scores of these groups to check that the hypothesised relationships hold.

Structural validity involves evaluating the dimensionality of the measure by assessing the underlying structure of the measure (Haywood, Garratt et al., 2005). Factor analysis, either confirmatory or exploratory, is used to identify components into which items group and to check for a dominant factor indicating that the measure is in fact only measuring the single concept it is intending to measure. Exploratory factor analysis (EFA) can be used to explore the number of potentially meaningful factors, representing distinct underlying concepts, within a measure while confirmatory factor analysis (CFA) can be used to examine the extent to which the data fit the predefined measurement model of a measure and explore whether there are alternative, more suitable factor structures which fit the data better. Criterion validity is very hard to establish in the validation of PROMs as there is no commonly accepted “gold standard” measure with which to compare (Fayers and Machin, 2016).

Content validity, as the degree to which the items of a PROM are an adequate reflection of the construct being measured, cannot be fully assessed using statistical methods. There are many different elements of content validity, including whether items are understood by respondents in the way developers intended, whether those items are relevant and acceptable to respondents and whether the measure comprehensively covers what the respondent considers to be important to that construct. Response rates can be investigated in order to see whether measures as a whole, or certain items, carry higher than expected missing response rates (Haywood, Garratt et al., 2005). Patterns in missing responses can indicate particular issues with acceptability. For example, higher missing response rates towards the end of the questionnaire can signal that it is too long while specific questions with higher non-response rates can signal that a question is either not understood, easily missed in the layout or considered irrelevant or inappropriate. However, researchers cannot be sure of the cause of the potential issues seen or the appropriate solution. This is because it is known that respondents will often still provide answers to

questions which they do not understand or do not feel are relevant or appropriate (Mallinson, 2002). This increases the likelihood of invalid responses, which may lead to misleading and biased results. Thorough examination of content validity requires the use of a qualitative study to investigate the interpretation of items in members of the patient or general population, whether they feel the items are relevant and appropriate and whether anything which is important to the construct of interest is missing from the measure. These methods will be described in more detail in section 2.5.2.3.

CTT reliability tests

Internal consistency reliability is commonly measured in CTT using Cronbach's alpha. Cronbach's alpha is a function of average inter-item correlation and the number of items in a scale, with values above 0.8 indicating good internal consistency but values over 0.9 argued to suggest item redundancy (Streiner, 2003). Cronbach's alpha coefficients are known to increase as the number of items in a scale increases. Therefore, this is not an unbiased measure of internal consistency (Fayers and Machin, 2016). Test-retest reliability is measured by estimating intraclass correlation coefficients between scores in the same participant over short periods of time, for example 2 weeks, where it is assumed no change in underlying health will have taken place (Fayers and Machin, 2016). Analysis of variance and intraclass correlation coefficients between scores can also be used to assess the inter-rater reliability of a person's score when a PROM is completed on behalf of the same individual, by two different raters at the same time (Fayers and Machin, 2016).

CTT responsiveness and sensitivity tests

The most commonly used tests of responsiveness and sensitivity are standardised response means (SRMs) and effect sizes (ESs). Sensitivity is investigated using cross-sectional comparisons of groups in which differences in QoL are expected. It is therefore closely related to known-group validity (Fayers and Machin, 2016), with the difference being that known-group validity looks to confirm that a difference in groups known to differ is shown by a measure, while sensitivity aims to show that clinically relevant or important differences between groups will be shown by the measure in a reasonably sized sample (Fayers and Machin, 2016). Responsiveness is concerned with a PROM's ability to detect within person change over time, where change has occurred using longitudinal data. One issue in the use of measures such as SRMs and ESs as measures of responsiveness and sensitivity is that these tests are based on the assumption that the underlying data follows a normal distribution (Fayers and

Machin, 2016). PROM response distributions are often non-normally distributed, and issues of floor and ceiling effects are fairly common. This means these tests may be biased (Fayers and Machin, 2016) and alternative, more robust tests may be required.

It is also possible to test responsiveness using criterion-based and construct-based methods (Mokkink, Terwee et al., 2010). Criterion based methods assess whether there is a statistically significant difference in the change scores of groups over time who do and do not meet the predefined external criteria, given a-priori hypotheses that these groups will have change scores of different directions or magnitudes. Construct based assessments of responsiveness compare the change scores between PROMs aiming to measure the same or similar constructs. Change scores can be compared by comparing mean differences of change scores or examining correlations between the change scores. Again, hypotheses about the expected direction and magnitude of correlations or mean differences of change scores should be made a-priori.

2.5.2.2 Item response theory and Rasch measurement theory

Item response theory (IRT) and Rasch measurement theory (Rasch) are the two main schools of thought in modern psychometric theory, both commonly used in psychometrics to develop measures and assess measure and item performance. Initially used in the building of parallel tests in education, these theories were born out of the observation that an individual's observed and true test scores are distinct from their "ability score", their true level of the underlying trait (Hambleton and Jones, 1993). A person's true or observed test score is test dependent and will therefore be lower on a more difficult test and higher on an easy test. However, their level of that trait (ability score) stays constant over all tests measuring that trait and is therefore test independent. This led to the desire to account for a respondent's amount of the underlying trait to enable superior examination of the measurement properties of items and tests and more precise estimation of scores.

The fundamental assumption of IRT and Rasch modelling is that a respondent with a given level of the latent unobservable trait, say QoL, will have a certain probability of responding in each response category of an item. This probability will depend on the "difficulty" of that question (Fayers and Machin, 2016), for example a question about someone's ease of walking 100m is "easier" to respond higher to than a question

about their ease of running 10km. Stochastic models, mostly variations on logistic item response models (Fayers and Machin, 2016), are used to estimate parameters representing the location of respondents and items on this latent QoL scale. Parameters are obtained by examining the probability of a specific item response as a function of the respondents' level of latent QoL and characteristics of the item (Chang and Reeve, 2005).

An important property of IRT and Rasch logistic models is that the exact value of the latent trait does not need to be estimated, as we are only interested in the relative difficulty of items and people's relative positions on the QoL scale (Fayers and Machin, 2016). Therefore both item parameters and respondent's level of latent QoL are expressed on the same relative scale (Fayers and Machin, 2016). These can be expressed on the log odds ratio (logit) scale ranging from about -4 to +4 (Fayers and Machin, 2016), although depending on the model type and parameterisation used, they can be scaled on the probit scale, which ranges from about -2.5 to +2.5. On both the logit and probit scale, zero represents the mean level of underlying QoL in the sample and non-zero values represent the number of standard deviations above or below the mean.

IRT and Rasch models describe item functioning using discrimination and difficulty parameters (Chang and Reeve, 2005). Discrimination parameters, one per item, examine how closely an item is related to the underlying QoL of respondents and how efficiently the item can discriminate between individuals with higher and lower QoL. Discrimination parameters can be an indicator of content validity, as item discrimination provides a measure of how closely related, and therefore relevant, included items are to the underlying QoL of respondents. An item with n response options also has $n-1$ difficulty parameters. The precise definition of difficulty parameters vary between different types of IRT model. As an example, in the Graded Response Model, each difficulty parameter tells us the amount of underlying QoL required to have a 50% probability of responding above a certain category, signifying better QoL, and a 50% chance of responding in that category or below. For example, given an item with response options 1-4, $b_1=-1$ tells us that someone 1 SD below mean QoL has a 50% chance of responding in category 1 and a 50% chance of responding above it. Therefore, difficulty parameters assess over what levels of QoL the item is able to precisely discriminate the QoL level of respondents.

There are also three-parameter IRT models, which include both these parameters and the pseudo-change (guessing) parameter, which assesses the likelihood that a respondent will guess the correct answer. However, these models are more relevant in educational testing rather than QoL measurement, where all questions are designed to be understood by all respondents, so guessing parameters are usually not used for psychometric analysis of QoL measures and will not be used in this thesis (Chang and Reeve, 2005).

There are a wide variety of IRT and Rasch models, which differ in the type of data they handle and the way it is summarised. Some models are built to handle dichotomous items, questions with two response options, while some handle polytomous items, for those questions with more response options. Models also differ in the assumptions they make about the ordering of response options. Ordinal models assume that there is a clear order in which the response options would be preferred, while nominal models are used for sets of response options where there is no clear ranking over which the options would be preferred. PROMs have a strong tendency towards response formats of more than two ordered response options per item and therefore this section will focus on polytomous ordinal IRT and Rasch models. A summary of the most commonly used polytomous IRT and Rasch models is shown in Table 1.

Table 1 – Characteristics of commonly used polytomous ordinal IRT models

Model	Item Response Format	No. of paras	Discrimination	Difficulty/threshold
Rating Scale Model (RSM)	Ordered categorical	1PL	Equal across items	Distance between category thresholds equal across items
Partial Credit Model (PCM)	Ordered categorical	1PL	Equal across items	Varies between items
Constrained GRM	Ordered categorical	1PL	Equal across items	Varies between items
Graded Response Model (GRM)	Ordered categorical	2PL	Varies between items	Varies between items
Generalised PCM	Ordered categorical	2PL	Varies between items	Varies between items
Nominal categories model	Nominal (does not force ordering of categories)	2PL	Varies between items	Varies between items

IRT and Rasch models also differ in the assumptions they make about their parameters. The Rasch family of models can be seen as special cases of IRT model. The Rasch family of models are commonly described as one-parameter IRT models (Chang and Reeve, 2005). One-parameter Rasch models allow difficulty parameters to vary between items but force discrimination parameters to be equal across all items, suggesting that all items are equally closely related to QoL. Two examples of one-parameter Rasch family models are the Partial Credit Model (PCM) and the Rating Scale Model (RSM).

The RSM handles ordered polytomous response options and requires that discrimination parameters are equal across items, meaning they are all equally well related to the trait (Chang and Reeve, 2005). It also assumes that the distance between item category threshold steps (like difficulty parameters) are equal across items and therefore the additional underlying trait required to be more likely to respond in category 3 rather than category 2 is the same across all items. These restrictive assumptions have led authors to question whether this model is appropriate for the type of data which stems from PROM responses (Nguyen, Han et al., 2014).

The PCM is another polytomous extension of the Rasch model (Hays, Morales et al., 2000). Again, it requires discrimination parameters to be equal across items but, unlike the RSM it allows the spacing between difficulty parameters and response categories to vary across items (Chang and Reeve, 2005).

Two-parameter IRT models allow both discrimination and difficulty parameters to vary between items. They do not force discrimination parameters to be equal across items and therefore allow for the fact that items may differ in how closely related they are to the underlying trait. Examples of commonly used two-parameter models are Samejima's Graded Response Model (GRM) and the Generalised Partial Credit Model (GPCM). Samejima's GRM is an extension of the two-parameter logistic model (Samejima, 1996). It assumes that item responses are ordered categorical (Hays, Morales et al., 2000). The GRM model allows both item discrimination and the spacing between each of the response categories to vary across items (Chang and Reeve, 2005). This allows for the fact that items may differ in how closely they are related to the underlying trait and that the amount of extra trait required to move between response options may not always be the same. A generalised two-parameter PCM (GPCM) has also been developed. The GPCM also allows the spacing between category thresholds to differ across items as well as discrimination parameters to

differ across items (Chang and Reeve, 2005). The nominal categories model, while not an ordinal model is also useful. As a nominal model, it does not assume a rank ordering of response options, allowing checks of whether they are being understood in the correct order

While the IRT and Rasch families of models are both based on logistic models and are therefore based on the same parameters and framework, there are important distinctions between the two which mean they have developed as separate theories, each with its own strong supporters (Fayers and Machin, 2016). Some authors have argued the use of two-parameter IRT models, such as the GRM model, because of its flexibility and the fact that it allows discrimination to vary item by item, meaning it typically fits response data better than the one-parameter Rasch family models (Reeve, Hays et al., 2007), in which discriminations are restricted to be equal across all items. Proponents of Rasch argue that the Rasch model is more robust and therefore selecting items which fit the Rasch model and rejecting those that don't is a better way of selecting well performing items (Fayers and Machin, 2016). However, it is claimed that this argument is more suitable to educational tests where there are infinite potential questions. This is not the case in PROMs, where item choice is limited to those which have face and content validity and it is therefore common that, to obtain reasonable fit, a two-parameter IRT model is required (Fayers and Machin, 2016).

The results of IRT can be displayed using item characteristic curves (ICCs) (Chang and Reeve, 2005). ICCs, one per item response option, express the probability of a respondent with a given level of QoL, selecting each response option of an item. Steeper ICCs indicate higher discrimination parameters and therefore items which are more closely related to QoL and more efficiently able to discriminate. Difficulty parameters determine the relative position of the ICCs on the QoL scale and therefore where on the scale the item can discriminate precisely the trait level of respondents. A well performing item is one with evenly spaced, steep ICCs covering a wide range of underlying QoL, with each response option, in the expected order, having a range of QoL over which it is the most likely option. ICCs can also give an indication of whether questions or item level labels are performing as expected.

Figure 2 shows the ICCs for two items. The x-axis shows the level of the underlying trait (θ) and the y-axis shows the probability of selecting each response option. The ICCs (coloured lines) of item 3 suggest that the response options are behaving well, as each option has a range of the trait over which it is the most likely response, and

the ordering of item response options reflects the severity of the levels. The ICCs are spread well across the range of θ , showing that the item is targeted to the full range of θ . By contrast, the ICCs of item 20 are clustered in the lower range of θ , which indicates that this item will be unable to distinguish the θ of respondents with higher levels of θ . This will result in substantial ceiling effects as anyone above average on the trait level will be most likely to respond at the ceiling for this item. Moreover, response option 2 is never the most likely option. This suggests an issue with performance of the item levels, as unordered or indistinct levels indicate that either the question or level labels are not being properly understood, or that the response options are indistinct from neighbouring categories.

Figure 2 - Examples of item characteristic curves and information curves for two items from a polytomous two-parameter Samejima's Graded Response IRT model.

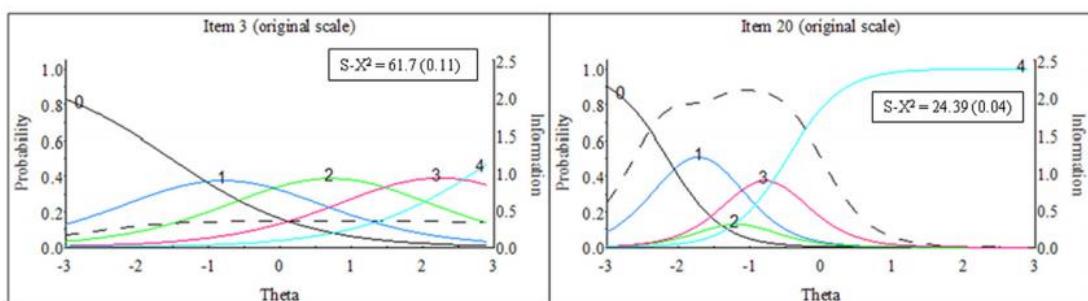


Figure 2 taken from Petrillo et al (2015) (Petrillo, Cano et al., 2015). Reprinted with permission by copyright holder Elsevier.

IRT examines the reliability and precision of measurement of an item or PROM in discriminating between individuals across the different levels of the underlying trait through item and total measure information levels (Chang and Reeve, 2005). Items which provide more information in a given area of the underlying trait contribute more to the overall precision of the test in that section of the underlying trait (Bjorner, Kosinski et al., 2003). The standard error of measurement (SEM) is inversely related to the information level ($SEM = 1/(information)$) (Cappelleri, Jason Lundy et al., 2014). Items with higher discrimination parameters provide more information and discriminate the QoL of respondents more precisely as they have a smaller item variance (Hays, Morales et al., 2000). The point on the underlying scale where item information is peaked is determined by the difficulty parameters. The total measure information level is the sum of item information. Total measure information can be

used to assess internal consistency reliability as total measure information is analogous to estimating Cronbach's alpha at each point on the QoL scale, with Cronbach's alpha=1-(1/total information) (Petrillo, Cano et al., 2015). Total information=5 equates to $\alpha=0.8$, a common cut-off for good internal reliability (Fayers and Machin, 2016).

Item or scale level information can be represented graphically using item or scale total information functions. The item information functions for two items are shown by the black dotted lines in Figure 2. Again, the x-axis shows the level of the underlying trait (θ). Reading from the alternative y-axis on the right-hand side of each graph, the y-axis shows the level of information provided by the item. The amount of information provided by the item is determined by the discrimination parameter. The spread of item information and the point on the underlying trait (θ) where this item information is peaked is determined by the difficulty parameters (Hays, Morales et al., 2000). The information curve for item 3 indicates that the item provides a fairly steady level of information across the full range of θ . The information curve for item 20 shows that it provides a lot of information in the lower half of the θ scale but very little in the top half. Therefore, item 3 is useful as a broad item for the full range of θ , while item 20 is good at precisely discriminating the underlying θ of respondents with poor QoL but is unable to discriminate with any precision those respondents with high QoL.

Differential Item Functioning

One of the key assumptions of the construct validity of PROMs is measurement invariance, which states that the PROM is measuring the same underlying construct in all groups of respondents. This means that relationships between the items and the latent trait; between the items themselves and between response options should be the same across groups. Therefore, an item about physical functioning should be equally important to the health of a 30-year-old as it is to a 75-year-old and the response category "moderate problems" on that physical functioning item should represent the same level of issue with physical functioning in those two individuals. DiF occurs when the measurement invariance assumption is violated and an item functions differently between subgroups of respondents (Fayers and Machin, 2007).

DiF arises because different groups systematically use or interpret response categories, or items themselves, differently (Knott, Lorgelly et al., 2017). Where DiF

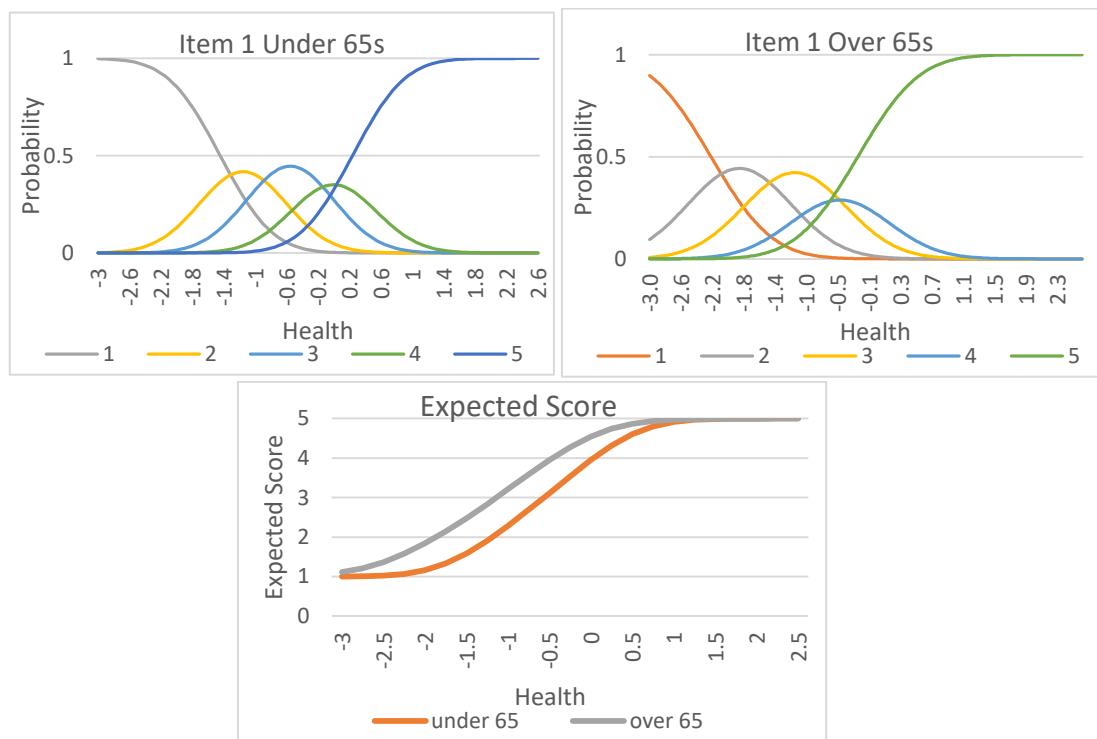
is present, respondents with the same level of QoL but who belong to different subgroups, have a different probability of providing the same response to the item (Chang and Reeve, 2005). Therefore, two individuals may rate their health differently, in part because their true level of health differs and in part because they interpret the categories differently (Knott, Lorgelly et al., 2017). This affects the overall measure score and therefore the estimate of QoL, threatening the measure's construct validity as it indicates that items are not solely measuring the construct they are hypothesised to measure but are also dependent on characteristics of respondents (Bjorner, Kosinski et al., 2003). Without an examination of DiF, we cannot know from simply looking at their resulting scores how much of this reported difference is true difference in health. If DiF is present, the estimate of the health, QoL or wellbeing of different groups will be biased by variables that are not the underlying attribute the scale intends to measure. This can in turn introduce bias into estimates of the effectiveness and incremental effectiveness of interventions and resulting ICERs on which resource allocation decisions are made in the economic evaluation of healthcare services.

Differences in item parameter estimates for different subgroups suggests the presence of DiF. DiF can be investigated in IRT frameworks by running multiple-group IRT models (Hays, Morales et al., 2000). These allow the same model to estimate the parameters of the relevant groups separately. A multiple-group IRT model where the parameters are forced to be the same in the two groups can be compared in terms of fit to a multiple-group model where the parameters are free to vary between groups. If the constrained model fits significantly worse than the model in which differences in item parameter estimates between groups are allowed, this suggests the presence of DiF between those groups (Hays, Morales et al., 2000).

The impact of DiF can be seen clearly through examination of ICCs and expected item scores. In the first panel of Figure 2, we saw the ICCs for an item, which tell us at any given level of the underlying trait the probability that a respondent will choose each response option. These probabilities can be used to calculate the expected score of respondents at each level of underlying trait on each item and the total measure. If we calculate the IRT parameters, ICCs and expected scores separately for different groups of respondents, we can see whether they differ and therefore whether the item exhibits DiF. For example, Figure 3 below shows the ICCs and expected scores for the same item in respondents aged 18-64 and 65+ separately. We can see from the ICCs that at the same level of underlying health, over 65s are more likely to respond in higher categories, signifying better health than over 65s. For

example, at an underlying level of health 2 SDs below the mean, over 65s are more likely to move away from category 1 and respond in category 2, signifying less problems to this item, while under 65s are far more likely to remain in category one, signifying more severe problems, until an underlying level of health approximately 1.4 SDs below the mean level of health. By this level of underlying health older adults are more likely to respond in level 3. Therefore, at each level of underlying health, until we reach the ceiling of the item, older adults are expected to respond higher to this item than younger adults, despite the fact that the two groups have the same underlying level of health. Therefore, this item exhibits DiF.

Figure 3 – ICCs and expected scores for under and over 65s for an item



IRT Model Assumptions

IRT models rely on three main assumptions: unidimensionality, local independence and monotonicity, which need to be checked before an IRT model is used. Unidimensionality means that the items of a measure should relate to a single latent trait (Chang and Reeve, 2005). As a consequence of that the respondent's level of the underlying trait accounts fully for the item responses (Cappelleri, Jason Lundy et al., 2014) and therefore for the item-variability in the scale (Fayers and Machin, 2007).

This can be investigated using factor analytic methods. PROMs may not be strictly unidimensional, but demonstration of the existence of a single dominant factor underlying the measure is sufficient (Reeve, Hays et al., 2007). However, PROMs, in their aim to cover all important aspects of broad concepts such as health and wellbeing, are often found to fail this assumption of unidimensionality. In response to this multidimensional IRT models have been developed to overcome this issue (Chang and Reeve, 2005). To be sure that the right type of model is run it is still important to thoroughly check the dimensionality of the measure.

Another important assumption of IRT models is local independence of items which states that there is no additional systematic covariance between items beyond their given relationship to the underlying trait being measured (Edelen and Reeve, 2007). Local dependence may arise in groups of items with similar content or which are physically grouped together on a measure. Local dependence may therefore signal item redundancy. Large modification indices of error covariances between items, or groups of similar items with substantially higher discrimination parameters than the rest of the items within a measure, may suggest Local dependence. One option is to remove one of the offending items at this stage, however this takes away the opportunity to gather additional information about the performance of this item. Longer scales, of approximately 10 or more items are also argued to be robust to Local dependence and therefore leaving these items in should not substantially impact models with more items (Edelen and Reeve, 2007).

The assumption of monotonicity requires that the probability of selecting an item response option which is indicative of a better health state increases as the underlying level of health increases (Reeve, Hays et al., 2007). This can be investigated by graphing item mean scores conditional on total score minus item score or examining the initial probability functions from non-parametric IRT models (Reeve, Hays et al., 2007).

There are no definitive rules regarding sample size requirements for IRT models. The more complex the model, the larger the suggested sample size, with two-parameter models requiring larger samples than Rasch type one-parameter models and multiple-group models requiring more respondents than single-group models. Sample sizes of 100 or more are said to be enough to estimate stable Rasch model parameters, while for more complex models with more parameters at least 500 are said to be required, with at least 200 to detect DIF (Edelen and Reeve, 2007). The

sample needs to well reflect the population of interest, across the full range of the measured construct continuum. It is necessary to have respondents endorsing all item response options for all items for the IRT model to be estimated (Edelen and Reeve, 2007).

2.5.2.3 Qualitative methods

Qualitative methods are the most appropriate way to assess content validity (Brod, Tesler et al., 2009) and have been widely used in the content validation of health, QoL and wellbeing measures in the literature (Clarke, Friede et al., 2011, Milte, Walker et al., 2014, Taggart, Friede et al., 2013, van Leeuwen, Bosmans et al., 2015b). Qualitative methods are required because the assessment of content validity involves investigating whether the questions being asked are an adequate reflection of the underlying construct being measured, in terms of whether they are understood by the respondents, acceptable to them, relevant and comprehensively cover what is important to their view of the construct being measured. This cannot be understood without qualitative methods which seek to directly capture the perspectives of respondents through methods such as interviews and focus groups (Brod, Tesler et al., 2009).

Qualitative methods aiming to ensure the content validity of a measure can be used at various stages in an instrument's lifetime. They are an important part of the measure development process, where interviews or focus groups with members of the target population of the measure can be used to generate information about important domains which should be covered within the measure based on participants' experience of either their condition or what is important to their health or QoL in general. Content validity of proposed versions of new measures or existing measures should also be assessed to ensure that the measure is being understood and performing as expected by developers, as well as whether there are any issues with the relevance or comprehensiveness of items. From the point of view of existing instruments, this is particularly important when there has not been a previous assessment of content validity in the specific population in which the measure is being used (Rothman, Burke et al., 2009). For example, the content validity of a measure may have been assessed in respondents aged 18-65, but this does not mean that the measure is understood/interpreted in the same way or has the same relevance or comprehensiveness in older adults.

Different qualitative methods are required at different stages in instrument development and validation. Semi-structured interviews and focus groups are used to generate relevant domains and items during measure development. This is important to attempt to develop a PROM which comprehensively captures everything that is important to the concept of interest. Once a proposed version of a measure has been developed, the content validity of this measure should be checked in terms of whether respondents understand the items as intended, whether those items are indeed relevant and acceptable to respondents and whether the resulting measure does in fact comprehensively cover what is important to the concept. Here the focus is not necessarily the generation of ideas and concepts, but it is to check whether the PROM performs as required when respondents complete it. Cognitive interviewing methods have been developed to explore the cognitive processes respondents go through when responding to a questionnaire and can therefore be useful to investigate any issues which arise during the completion of the measure (Willis, 2005). As this thesis focuses on existing PROMs, it will focus on content validation methods relevant to proposed and established measures, rather than earlier stages of measure development.

Cognitive interviewing techniques are often used to assess content validity as they enable the researcher to explore the thought processes respondents go through when answering survey questions and the factors which influence the answers provided (Collins, 2003). Cognitive methods are based on theories of survey response, such as the question-and-answer model developed by Tourangeau (Tourangeau, 1984). The question and answer model details four stages that respondents go through when answering a survey question; comprehension, retrieval, judgement and response. First the respondent has to understand the question (comprehension), then they must retrieve the valid information from their memory (retrieval), make a judgement about the information needed to answer the question (judgement) and lastly, they must choose a response to the question (respond).

There are a variety of points in this process where issues may arise which threaten the validity of responses. In the comprehension phase, respondents may not understand the question or response options, or may interpret them differently to how the measure developers intended. It is important not only that respondents understand the question, but that respondents interpret both the question and response options in the same way as the developer intended, otherwise conclusions made based on the answers given may be flawed and comparisons between different

respondents' answers will not be valid (Collins, 2003) as they may essentially be answering different questions. In the retrieval phase, there are a number of potential issues which can occur when recalling information. The respondent may be unable to recall the information because: it never reached long term memory; they cannot distinguish it from similar information or events; or it may have been tainted as the respondent struggles to remember the exact details and so attempts to fill in the blanks with information which may be inaccurate (Collins, 2003).

In the judgement phase the respondent goes through several processes to formulate their answer (Collins, 2003), including assessing whether they understand the question, whether it applies to them, whether they have the information required to answer, how detailed and accurate their answer needs to be and whether they should modify their answer to meet the perceived needs of the question. They may employ judgement heuristics, which are cognitive shortcuts used to estimate answers where they feel they do not have complete recall, where they feel the accurate information is too difficult to reach or they simply feel that the questions asked are not relevant or appropriate to them and may not engage fully with the question. In the response stage there are two important components, each with potential to create response errors. First the respondent goes through the process of response formatting where the respondent must fit their personal answer into one of the response options provided (Collins, 2003). Here there may be a mismatch between the options provided and the desired response of the respondent or the response options provided may affect the way the respondent answers as they may suggest a "usual" behaviour which the respondent feels they should adhere to. This leads onto the final possible issue; response editing, where the respondent may feel social pressure to respond more positively than their true state (Collins, 2003).

Cognitive interviewing methods were developed by Willis, with the aim of exploring peoples' response processes in an attempt to pinpoint when and which types of response issues occur (Willis, 2005). Two commonly used cognitive interviewing techniques are think-aloud and verbal probing. Think-aloud techniques ask respondents to verbalise their thought processes as they complete a questionnaire. Verbal probing involves asking respondents specific questions in order to understand how they arrived at their chosen response either during the completion of a questionnaire (concurrent verbal probing) or after questionnaire completion (retrospective verbal probing) (Collins, 2003). Different probes can be used to explore different stages of the response process, for example a respondent's comprehension

of the question can be explored using the probe “What does X mean to you?” or the response stage can be explored by asking “Were you able to find your answer to the question in the response options shown?” (Collins, 2003). These techniques will be discussed further in Chapter 5.

2.6 Conclusion

This chapter outlined the current methods of economic evaluation as a means of making resource allocation decisions in the NHS in the UK, with a focus on how the effectiveness of health interventions is measured in CUA using QALYs. We examined NICE’s current preference for measuring the QoL element of the QALY using the EQ-5D measure of HRQoL in healthcare evaluation. The current policy interest in extending these methods to evaluate both social care and integrated health and social care services were then outlined. This included consideration of adaptations which may be required to current outcome measurement practice, in order to appropriately and comprehensively measure the outcomes of social care interventions, which often fall outside of health improvement. The argument for extending the QALY beyond health to include broader elements of QoL and wellbeing was put forward, before the definitions and conceptualisations of these constructs in the field of outcome measurement were considered.

Then the process required to create a well performing PROM of any of these concepts was described. Lastly, the elements of psychometric performance which should be checked to ensure that a measure is a valid, reliable and responsive measures of the intended construct in those people in which it will be used were then outlined as well as the different methods available to investigate the performance of measures.

Chapter 3

Selection of existing PROMs potentially suitable for evaluating health and social care interventions and investigation of their psychometric performance in older adults

3.1 Introduction

There are a huge range of generic PROMs available for use in the literature and it would not be possible to assess the psychometric performance in older adults of all potentially relevant PROMs in a single piece of research, such as this one. Therefore, before work could begin, a list of potentially relevant generic PROMs needed to be identified. The methods used to identify potentially relevant generic PROMs and the justifications for the resulting choice of measures included in this thesis are outlined in this chapter. Then the process by which the chosen PROMs were developed and tested is outlined. Next a systematic review investigating the psychometric evidence for these PROMs in measuring the health, QoL or wellbeing of older adults is presented, before the rationale for the research, in the context of research gaps identified in the systematic review is outlined. Finally, the aims and objectives of the study are presented, and the study design chosen to meet these aims and objectives is introduced.

3.2 Identification of relevant PROMs

3.2.1 Methods for identifying relevant PROMs

Several methods were used to identify PROMs which would be potentially relevant to evaluate health and social care interventions in older adults. First, relevant policy documents, such as the NICE health and social care guidelines (National Institute for Health and Care Excellence, 2013, National Institute for Health and Care Excellence,

2016) were searched to understand the current approaches and suggestions for outcome measurement in the field. Since social care funding decisions are largely made by local government, who may not solely rely on NICE guidance, broader policy documents relating to outcome measurement for evaluating public policies in relation to health and social care were also searched for. Next, a rapid review of PROMs which have been used to evaluate integrated health and social care interventions was conducted in order to identify which measures were being used in practice. Lastly, expert opinion was sought from local researchers involved in evaluations of health and social care services aimed at older adults, in order to investigate which measures they had used and their experience of them.

3.2.2 NICE and policy documents

NICE have published separate guidance manuals for the evaluation of healthcare and social care interventions (National Institute for Health and Care Excellence, 2013, National Institute for Health and Care Excellence, 2016). NICE's remit for the evaluation of healthcare interventions is much more established and, in this guidance, they are clear that CUA evaluation should be performed, with the QALY used as the unit of effectiveness. They state that the EQ-5D should be used to assess the impact of interventions on the QoL of patients and to generate utility values for use in QALY calculations, unless the EQ-5D has been shown to be inappropriate in that population or condition.

However, NICE's remit over the evaluation of social care interventions is much more recent, and CUA evaluation methods and outcome measurement in this field is much less developed than in healthcare. Therefore, while NICE state that they would prefer a CUA, which allows effective comparisons across different social care interventions, they acknowledge that this may not currently be possible as there is currently no accepted social care equivalent to the healthcare QALY (National Institute for Health and Care Excellence, 2016). The social care guidance manual states that economic evaluations in social care should measure and value effects using; the healthcare QALY based on the EQ-5D where health effects are relevant, or a "social care" QALY, based on the ASCOT, with a parallel evaluation based on a capability (ICECAP) or wellbeing measure where an intervention has outcomes in health, social care and capabilities. The choice of wellbeing measure is left open. This flexible approach to

the measurement of outcomes has issues for the comparability of economic evaluations and resource allocation in social care.

Beyond NICE, there have been other programmes which have aimed to develop outcome frameworks for assessing the outcomes of public policies. The Office for National Statistics (ONS) has developed the Measuring National Wellbeing Programme which, having emerged outside of traditional health economics, involves a wider set of domains and indicators than we would expect from measures within health economics. The ONS Measuring National Wellbeing Programme consists of 10 domains each with 3-5 indicators (Oguz, Merad et al., 2013). These domains include; health, relationships, personal wellbeing, what we do, where we live, personal finance, the economy, education and skills, governance and the national environment. The personal wellbeing domain has been used as a measure of wellbeing in its own right, in attempts to measure personal wellbeing in evaluations of health and social care interventions. It includes four questions, treated as separate indicators of different aspects of personal wellbeing, which have together become known as the ONS-4 (Office for National Statistics, 2016c). The ONS-4 items cover; life satisfaction, the extent to which people feel the things they do in their life are worthwhile, happiness and anxiety; each with an 11-point response scale measured from 0 “not at all” to 10 “completely”.

The Measuring National Wellbeing Programme has also included the Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS) (Peasgood, Brazier et al., 2014, Stewart-Brown, Tennant et al., 2009), a seven-item positive mental wellbeing measure developed from the original 14 item WEMWBS. The WEMWBS and SWEMWBS have also been included in several large population surveys including Understanding Society and the Health Survey for England (Health and Social Care Information Centre and Department of Health, 2014), which seek to monitor population health and wellbeing. The What Works for Wellbeing group also mention the WEMWBS measures and the ONS-4 in their work towards encouraging decision makers to evaluate public policy interventions based on high quality wellbeing evidence (What Works for Wellbeing, 2018). This shows a broader interest in the use of wellbeing measures in the evaluation of public policy interventions.

3.2.3 Rapid review of PROMs which have been used in evaluations of integrated health and social care interventions

3.2.3.1 Introduction

The aim of this section is to identify those generic PROMs which may be suitable for the evaluation of health and social care interventions. It was decided that one way to do this was to identify the generic PROMs which have been used to evaluate the outcomes of integrated health and social care interventions. The researcher chose to focus on integrated health and social care interventions both because of the recent policy focus on integration between health and social care services and because it would be expected that evaluations of integration would aim to capture the relevant outcomes to both sectors. If a PROM is being used to evaluate an intervention with input from, and outcomes relevant to both health and social care, the evaluators must feel that that PROM is able to comprehensively capture the impact, in terms of both health and social care outcomes, of that intervention on participants.

There are many definitions of integration, depending on the context, the extent of integration and the organisations involved (Robertson, 2011). It is generally described as the bringing together of inputs, delivery, management and the organization of services in order to improve access, quality, user experience and efficiency (Kodner and Spreeuwenberg, 2002). Integration between health and social care could involve the joint delivery of services, joint commissioning and/or funding of services across multiple organisations, or, taken further, the organisations themselves could be integrated, with a single organization delivering both health and social care (Robertson, 2011). Regardless of the depth of integration between these two sectors, an impact on the outcomes of both sectors would be expected and should be evaluated.

3.2.3.2 Methods

Electronic databases were searched using a search strategy combining a variety of terms for QoL, integration and evaluation in health and social care, detailed in Appendix 1. These terms were searched for in both titles and abstracts. This database search covered MEDLINE, CINAHL and Social Care Online from their inception to April 2016. These databases were chosen because they were thought to adequately span the health and social care literature. In addition, the references of included

papers and any relevant reviews were hand searched for further potentially relevant papers. Inclusion and exclusion criteria were as follows.

Inclusion criteria

- Evaluations of integrated health and social care services or interventions
- Studies had to be available online and in English

Exclusion criteria

- Studies that were not explicit that the service or intervention included both health and social care
- Studies which failed to include a generic PROM measuring health, QoL or wellbeing
- Studies investigating integration in childrens' services were excluded as this study sought PROMs suitable for adults
- Policy and descriptive documents which did not include an evaluation of integrated services
- Reviews themselves were not included however the included papers of any relevant reviews returned were retrieved and considered for inclusion

As discussed in the introduction of this section, integration can range from joint working between health and social care on small individual projects to pooled funding to fully integrated organisations. The extent of integration is often not made clear in the reporting. Therefore, for the purpose of this review any explicit mention of joint working between health and social care, such as multidisciplinary teams, was considered sufficient to be classed as integration, even if it were only for small groups of patients or at a local level. This is sufficient because the assumption that the outcome measures used should be relevant to all health and social care outcomes of the intervention still holds.

To be included in the review, QoL had to be measured using standardised PROMs. These PROMs could measure health, QoL, SCRQoL or wellbeing. For the purposes of cross-sectoral evaluation, it was decided that generic measures of QoL were of interest in order to be able to compare across disease areas. Therefore, condition-specific measures found in accepted full text papers were excluded. Patient reported measures that simply screened for symptoms or conditions were also excluded. For

a generic measure to be included the entire PROM had to have been used and not altered. Use of sections of PROMs, or of individual questions from a measure did not count as full use as this would not generate comparable results between evaluations. From those studies accepted at full text all potentially relevant PROMs were extracted. At this stage, it was not always clear whether these measures were QoL measures or clinical screening and symptom measures. Therefore, once the full list of measures was extracted each measure was investigated further and a classification system for the measures was developed by the author.

Once measures had been judged to be generic QoL measures they were further classified as measuring health, SCRQoL, capabilities or wellbeing according to the measure description given by the developers and definitions of concepts identified in section 2.3. There were some issues identified in this stage. For example, several mental health measures were returned which covered only a single concept, such as morale in the Philadelphia Geriatric Centre Morale Scale. Although this is a concept that could be considered as an element in an individual's wellbeing, a measure including only this domain was not considered wide enough to fully capture wellbeing, as a wellbeing measure in its own right and therefore was classified as a mental health condition-specific questionnaire rather than a generic wellbeing measure. The 12-item general health questionnaire (GHQ- 12) was also difficult to classify. It was developed as a screening tool for minor psychiatric disorders however, it does contain questions that could be defined as wellbeing. While it has been used as a wellbeing measure in the literature where other measures were not available, it was not developed for this purpose and has been argued not to be appropriate for making inter-personal or inter-temporal comparisons, since the response options refer to whether the individual feels better or worse than usual (Alshreef and Dixon, 2015). Therefore, again it was excluded as a generic wellbeing measure. Similarly, several activities of daily living measures were returned. While often fairly generic these measures were considered a measure of daily functioning, which could be argued to be an aspect of QoL, rather than a measure of QoL itself and were also excluded. Generic QoL measures were then further classified as preference and non-preference-based.

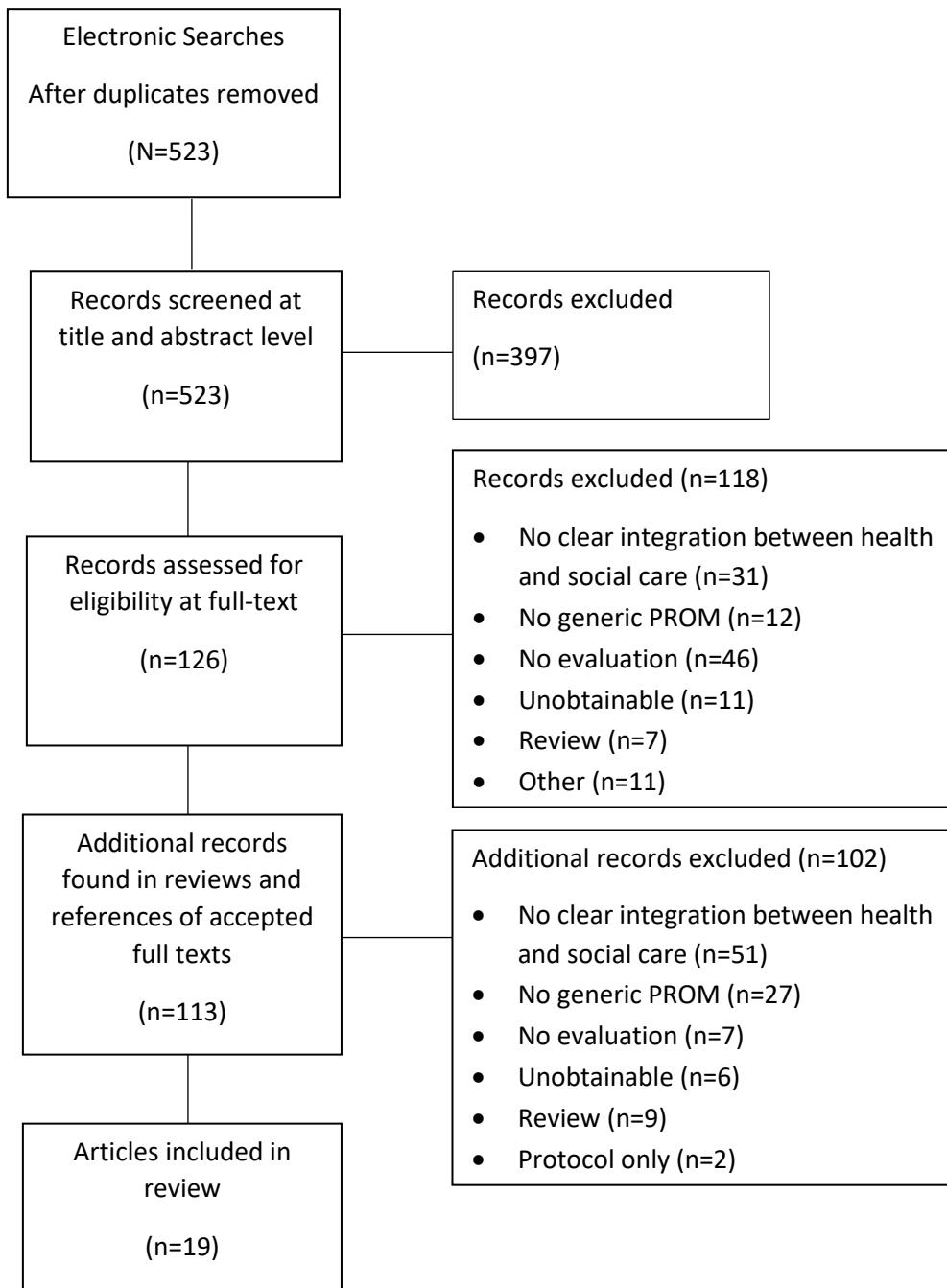
Data was extracted from included papers on the intervention, setting, study design, study population and QoL measures used. Quality appraisal was not carried out as the aim of this review was only to identify potentially relevant PROMs and therefore study design and methodological quality were outside of the remit of this review.

3.2.3.3 Results

The PRISMA diagram is displayed in Figure 4. The electronic database search returned 629 records. After duplicates were removed this was reduced to 523 records to be screened at title and abstract level. All screening was completed by the author. Three hundred and ninety-seven records were excluded as irrelevant using the inclusion/ exclusion criteria at this stage, leaving 126 records to be assessed at full text. One hundred and six records were excluded at this stage due to; no clear integration between health and social care (n=32), failure to include a generic PROM (n=12), the record was descriptive and did not give the results of an evaluation (n=46), the record was a review (n=7), study population was children (n=3), protocol only (n=6), non-English language (n=1), abstract only (n=1) and record was unobtainable (n=11). From the references of the eight included papers and the relevant reviews found in the electronic search, another 113 records were assessed for eligibility. From these 75 records were excluded due to no clear evidence of integration between health and social care (n=51), failure to include a generic PROM (n=27), not an evaluation (n=7), review (n=9), protocol only (n=2) and unobtainable (n=5). All study results of protocols were either captured elsewhere in the search (n=4), or results were not yet published (n=4) according to ClinicalTrials.gov (U.S National Library of Medicine, 1993-2018) and the ISRCTN registry (BioMed Central Ltd, 2019).

The search resulted in 19 papers describing 17 integration schemes which included a generic PROM, being accepted at full text. From these, seven different PROMs were identified. Five of the generic measures identified were judged to be measuring health: the EQ-5D (Brooks and The EuroQol, 1996), the Nottingham Health Profile (NHP) (Hunt, McKenna et al., 1980) and the 36 item, 20 item and 12 item Short Form Health surveys (SF-36, SF-20 and SF-12) (Maruish, 2012, Ware, Snow et al., 1993). Of the remaining two identified PROMs, one measured SCRQoL (ASCOT) (Netten, Burge et al., 2012) and one capabilities (ICECAP-O) (Grewal, Lewis et al., 2006). Three of the identified measures were preference-based; the EQ-5D, ASCOT and ICECAP-O, while the remaining 4; the Nottingham Health Profile (NHP) and the 36 item, 20 item and 12 item Short Form Health surveys (SF-36, SF-20 and SF-12) were non-preference-based. However, it is worth acknowledging that it is possible to derive a preference-based score from the SF measures via responses to certain items that form the preference-based SF-6D instrument (Brazier and Roberts, 2004).

Figure 4 - PRISMA Diagram



Appendix 2 shows the 17 integration schemes identified, along with the generic QoL measures collected in each. Table 2 shows the PROMs identified from these schemes, what they measure and the number of times they were found. Eleven of the 17 studies (65%) identified used the EQ-5D (Ariss, Enderby et al., 2015, Cartwright, Hirani et al., 2013, Gage, Grainger et al., 2014, Hammar, Rissanen et al., 2009, Henderson, Knapp et al., 2013, Hultberg, Lönnroth et al., 2005, Hultberg, Lönnroth et al., 2007, Jones, Forder et al., 2013, PricewaterhouseCoopers and Australian

Government Department of Health and Ageing, 2007, Reid, 2007, Sahota , Pulikottil-Jacob et al., 2016, Sulch, Melbourn et al., 2002, Sulch, Perez et al., 2000, Windle, 2009), while six of the 17 (35%) used the SF-36 (Anderson, Mhurchu et al., 2000, Gage, Grainger et al., 2014, Harris, Ashton et al., 2005, Lumley, Watson et al., 2006, PricewaterhouseCoopers and Australian Government Department of Health and Ageing, 2007, Sommers, Marton et al., 2000). The Nottingham Health Profile was included in two studies (Anderson, Mhurchu et al., 2000, Hammar, Rissanen et al., 2009), while one study each included the ICECAP-O (Cartwright, Hirani et al., 2013, Henderson, Knapp et al., 2013), ASCOT (Jones, Forder et al., 2013), SF-12 (Cartwright, Hirani et al., 2013, Henderson, Knapp et al., 2013) and SF-20 (Toseland, O'Donnell et al., 1996). Five schemes gathered data on more than one measure of generic QoL, of which four included the EQ-5D, in combination with; SF-12 and ICECAP-O; NHP; ASCOT and SF-36. The only other measures collected together in a scheme were the SF-36 and NHP. All studies which included the EQ-5D, except one (Jones, Forder et al., 2013), provided enough detail to establish that the 3L version was used. The remaining EQ-5D study (Jones, Forder et al., 2013) was published in 2013, two years after EQ-5D-5L was first published in 2011. Therefore, it is most likely to have used the 3L.

Table 2 – Generic QoL measures identified

Generic Measure	Health	SCRQoL	Capability	Wellbeing	Pref-based	Non-pref based	No. studies used
EQ-5D	1				1		11
SF-36	1					1	6
ICECAP-O			1		1		1
ASCOT		1			1		1
NHP	1					1	2
SF-12	1					1	1
SF-20	1					1	1

The EQ-5D-3L preference-based measure of health status (Brooks and The EuroQol, 1996) was described in section 2.2.1. Details of the other generic measures identified are outlined below.

The ICEpop CAPability measure for older people (ICECAP-O) (Coast, Flynn et al., 2008, Flynn, 2011) is a measure of capability designed for use in economic

evaluations aimed at elderly populations. It focuses on wellbeing with five attributes; attachment, security, role, enjoyment and control. Each attribute has four levels ranging from no capability to full capability. It is preference-based; however, the anchors are 0=no capability and 1=full capability. This means that currently it cannot be used to produce a QALY utility value comparable to the health QALY as the anchors differ from the required death and full health. Although it wasn't used in any of the studies we uncovered, there is also an ICECAP-A, designed for an adult rather than elderly population (Al-Janabi, Flynn et al., 2012). It contains five attributes; attachment, stability, achievement, enjoyment and autonomy. Again, each attribute has four levels ranging from no capability to full capability. It is preference-based on the same anchors to the ICECAP-O and therefore also cannot be used in the health QALY model as it is currently valued.

The ASCOT is a measure of SCRQoL, which aims to measure the extent to which individuals' social care needs and wants are being met. It has eight domains; control over daily life, personal cleanliness and comfort, food and drink, accommodation cleanliness and comfort, safety, social participation, occupation and dignity. Each domain contains one item with four response levels; ideal, no unmet needs, some unmet needs and high unmet needs. ASCOT is preference-based with 0 equivalent to death and 1 representing full SCRQoL where all social care wants and needs are met. There are two sets of possible scores. One from a general population valuation study using best worst scaling (BWS) and another which has used TTO values for a sample of states to anchor the BWS scores onto the QALY scale with 0 equivalent to dead (Netten, Burge et al., 2012).

The Nottingham Health Profile (NHP) (Hunt, McKenna et al., 1980, Hunt, McKenna et al., 1981) is a measure of subjective health status, which aims to determine the effect a disease has on QoL. It has two parts. The first is a 38-item health questionnaire which focuses on six areas; pain, energy, sleep, mobility, emotional reaction and social isolation. The second part, which can be omitted, includes seven items which focus on the areas of life affected including; occupation, housework, social life, family life, sexual function, hobbies and holidays. It is non-preference-based, with scores simply summed, with higher scores indicating worse health.

The 36-item short form health survey (SF-36) (Ware, Snow et al., 1993) is a measure of health status which covers eight domains; physical functioning, role limitations due to physical problems, pain, general health, vitality, social functioning, role limitations

due to emotional problems and mental health. Scores are summed, with higher scores indicating better health. Two composite scores can also be calculated for physical and mental health. The SF-20 (RAND, 2015b) and SF-12 (RAND, 2015a) are shorter 20 and 12 item versions of the SF-36 which are also non-preference-based. However, preference-based scores can be generated by mapping responses to the SF-6D preference-based measure of health.

3.2.3.4 Discussion of rapid review findings

The EQ-5D and SF health measures were by far the most commonly used generic PROMs in evaluations of integrated health and social care services. This is unsurprising, as these measures are well established in the health sector and the EQ-5D is the measure required by NICE for the evaluation of healthcare interventions. It is promising to see that the three preference-based measures returned in the literature review; the EQ-5D, ASCOT and ICECAP-O, are those which have been suggested by NICE for use in evaluations of social care in their social care guidance manuals (National Institute for Health and Care Excellence, 2016, National Institute for Health and Clinical Excellence, 2013). The fact that ASCOT and ICECAP-O were only used once each may be due to the fact that they are much newer than the more established preference-based EQ-5D and non-preference-based SF-36. They have both fairly recently become preference-based, which could mean their use will escalate, given time. The ASCOT and ICECAP-O were both used in combination with the EQ-5D, potentially due to recognition that they may include broader elements of QoL, which may be relevant to integrated health and social care evaluations but missed by the EQ-5D as a measure of health. The NHP was found to have been used in two studies. However, this is one of the older and longer measures identified and its use has declined in recent years, potentially due to the availability of shorter preference-based measures.

There are several important limitations to this rapid review. A limited number of databases were searched. These databases were chosen because they were thought to best cover the breadth of health and social care literature. It is hoped that the databases chosen in combination with searching references from included papers and included studies of identified reviews, will have sufficiently covered the available literature, however it has to be acknowledged that some studies may have been missed. Seventeen records were also unobtainable, either because they were

unavailable for free online or by interlibrary loan request or they were policy documents or reports that could not be found on current versions of websites.

3.2.4 Researcher experience

Researchers involved in a range of local evaluations of integrated health and social care projects were consulted to investigate which PROMs they were using and their experience with these. A local Sheffield based BCF integrated health and social care project called People Keeping Well in their community was evaluated from a health and social care perspective within ScHARR. People Keeping Well was a community-based social prescribing prevention intervention which aimed to prevent and delay health and social care service use (Sheffield Clinical Commissioning Group, 2015). This project collected three PROMs; the EQ-5D, ONS-4 and SWEMWBS. However, only the EQ-5D was included in the evaluation. An evaluation of a BCF social prescribing intervention in Doncaster also used the EQ-5D-3L as well as a life satisfaction question, similar to that included in the ONS-4. A Vanguard project based in care homes in Wakefield (Healthwatch Wakefield, 2016) also included the ONS-4.

These projects reported a range of experiences with these measures. Care home participants were reported to have experienced negative emotional responses to ONS-4 questions in the Vanguard project in Wakefield (Healthwatch Wakefield, 2016). This caused the project team to remove the ONS-4 worthwhile question as it caused pilot interviews to “take an unhelpfully negative trend” (Healthwatch Wakefield, 2016). The People Keeping Well project in Sheffield also reported anecdotally that participants had issues with the SWEMWBS and ONS-4. In the Doncaster evaluation, while participants had no problem with the life satisfaction question they did not like the EQ-5D.

These experiences were mostly anecdotal and often relayed from interviewers to the researchers evaluating the intervention. Researchers noted that it was difficult to be sure in some cases whether it was participants who reported issues with the measures or whether the issue was that interviewers felt that the measures were inappropriate, but instead reported that participants were not comfortable with the questions asked. Therefore, they felt it was important to investigate the content validity

of the EQ-5D, WEMWBS and ONS-4 directly in populations which are commonly targeted by these types of interventions.

3.3 Selection of PROMs for inclusion

Having identified a range of potential PROMs which could be included in this PhD research, the decision of which PROMs to include was based on several factors:

- i) Measures commonly or locally used to assess QoL in health and social care evaluations
- ii) Measures recommended for possible use by NICE in their health and social care guidelines or by broader outcome measurement policy documents being used to inform policy making in these sectors
- iii) Generic measures potentially suitable across a range of populations and conditions
- iv) Availability of data for conducting psychometric analysis of measures

The measures chosen and the justifications for their inclusion are outlined in the following section.

EQ-5D

The EQ-5D (Brooks and The EuroQol, 1996) preference-based measure of health was selected as an important PROM, as it is the most commonly used measure in health in the UK. It is required by NICE in the CUA evaluation of healthcare interventions in England (National Institute for Health and Care Excellence, 2013) and is also mentioned in the NICE social care guidance as a suggested measure of benefit in evaluation of social care interventions (National Institute for Health and Care Excellence, 2016). However, it has been suggested that this measure is inappropriate in social care evaluation as it fails to account for broader non-health aspects of QoL (Brazier and Tsuchiya, 2015), such as dignity and control, which may limit its usefulness in cross-sectoral evaluation.

ASCOT

ASCOT has been developed as a preference-based measure of SCRQoL, specifically for the evaluation of social care services (Netten, Burge et al., 2012). ASCOT is one of the PROMs recommended by NICE for use in evaluation in their social care evaluation guidance (National Institute for Health and Care Excellence, 2016). The ASCOT is much newer than the EQ-5D and it is becoming increasingly popular. As the only QoL measure found in the rapid review which is specifically aimed at social care users, it was thought important to examine its psychometric performance in an older population. ASCOT should be better tailored to aspects of QoL where social care services are likely to have an impact, than health measures such as EQ-5D, as its focus lies on an individual's practical capability to function day to day.

WEMWBS

WEMWBS was developed as a measure of positive mental wellbeing in the general population (Tennant, Hiller et al., 2007). There is also a reduced version, the SWEMWBS that consists of seven of the fourteen questions from the full WEMWBS (Stewart-Brown, Tennant et al., 2009). The SWEMWBS has been included in the Office of National Statistics (ONS) Measuring National Wellbeing Programme (Peasgood, Brazier et al., 2014). Wellbeing measures are also mentioned in the NICE social care guidance as potentially appropriate in the evaluation of social care interventions. Although the WEMWBS is not preference-based it is one of the more popular wellbeing measures. Another important reason for the inclusion of the SWEMWBS was that this, alongside the ONS-4, was one of the measures to which participants had negative reactions in the Sheffield People Keeping Well BCF Project. Therefore, checking its validity and appropriateness in an older population was considered of key importance.

ONS-4

In 2011 the ONS began to measure personal wellbeing using the ONS-4 as part of the ONS Measuring Subjective Wellbeing Programme. The ONS-4 has also been recommended to decision makers looking to evaluate public policy interventions by the What Works for Wellbeing Centre (What Works for Wellbeing, 2018) and suggested as potentially relevant in the NICE social care guidance (National Institute

for Health and Care Excellence, 2016). While the ONS-4 was not developed as a preference-based measure of QoL or wellbeing suitable for economic evaluation, use of the ONS-4 in the health and social care sectors as a measure of QoL or wellbeing has been increasing. There has been a particular push for their use in projects with local government involvement. Therefore, it is important to consider its validity as a measure of wellbeing.

There have been issues in the collection of ONS-4 data in local wellbeing data collections. Patients have reacted badly to ONS-4 questions in data collection interviews in a Vanguard project in Wakefield (Healthwatch Wakefield, 2016) and a BCF project in Sheffield. In the Wakefield project this caused one of the ONS-4 questions to be removed as it caused distress in pilot interviews (Healthwatch Wakefield, 2016). This, combined with the lack of reporting on the validity of ONS-4 questions in older adults, has highlighted the need to investigate this measure further.

SF-12v2

The SF instruments, particularly the SF-36, were identified as among the most popular PROMs in the evaluation of integrated health and social care services in the rapid review. When this project was discussed with experts in the care of older and frail individuals, who were involved in running the Community Ageing Research 75+ Study (CARE 75+) (National Institute for Health Research, 2014), a UK cohort study of health transitions and frailty in those aged over 75, they revealed that they see the SF-36 as an important measure in this population as it includes many aspects of QoL that they consider important in an older frail population including social inclusion. Therefore, the SF-12v2 measure was included as it is short enough to be feasibly included, while retaining a good coverage of the full SF-36 questions.

Other measures considered

Other measures were considered for inclusion in this thesis. The NHP was identified in the rapid review as having been used in two evaluations of integrated health and social care services. However, the NHP is one of the longer and older measures which have been identified and its use as a health measure has declined in recent years,

with the EQ-5D and SF measures proving much more popular, possibly due to their relative brevity and options for preference-based assessment. Therefore, the NHP was not chosen for inclusion in this study.

The ICECAP measures of capability (Al-Janabi, Flynn et al., 2012, Coast, Flynn et al., 2008, Grewal, Lewis et al., 2006) were also suggested by the NICE social care guidelines as potentially relevant measures for the evaluation of social care interventions (National Institute for Health and Care Excellence, 2016). These measures were not included for several reasons. There are two different versions of the ICECAP instruments; the ICECAP-A which is aimed at adults aged 18-64 and the ICECAP-O, which is aimed at older adults (aged 65+). These different versions include different items. This caused issues in several aspects of this research. Firstly, a direct comparison of the psychometric performance of these measures in older and younger adults would not be possible as they are in fact different measures. There could not have been an examination of DiF, which forms an important part of the first study in this thesis. Secondly, this thesis began with the aim of investigating generic measures, which are potentially suitable for examining the health, QoL or wellbeing of all respondents as, to ensure comparability between economic evaluations, the same measure must be used. This would not be the case using the ICECAP measures, as separate measures would be used for under and over 65s.

By including measures such as the ONS-4 and SWEMWBS in the National Wellbeing Programme the government is recommending their use in evaluation, public policy making and resource allocation. NICE does this for EQ-5D and ASCOT by including them in its guidance for health and social care evaluation, and directly recommends the use of wellbeing measures, of which the WEMWBS and ONS-4 are two popular options. Additionally, all of the above measures were being used locally in various studies involving older adults and therefore it was felt important to check the validity of these measures in this important health and social care population. The existing evidence on the validity of each of the included instruments is investigated in a systematic review in section 3.5. However, before we can investigate the validity of the included measures it is important to first understand how they were developed. This will be outlined in the next section.

3.4 Development of included PROMs

Details of how each measure was developed and tested by their development team are outlined in this section. Particular attention is paid to whether patients or members of the public were involved in the development of the measure as this can provide important support for the content validity of the measure. Examination of the psychometric testing carried out by the development team is also an important starting point from which evidence is built on the broader psychometric performance of the measure. Particular attention will be paid to whether measure development and psychometric testing included the input of older adults, as this population is often overlooked in testing, despite being an important group of health and social care users.

3.4.1 EQ-5D

The EQ-5D, at first called the EuroQol instrument, was first released publicly in 1990 (Brooks and The EuroQol, 1996). It was developed by the EuroQoL group, a multi-disciplinary international group, with the aim of creating a generic PROM to measure and value health status (Devlin and Brooks, 2017). The generic descriptive system was developed by performing a detailed review of existing generic health measures and using expert judgement within the group to select and refine possible domains (Devlin and Brooks, 2017). The EuroQol instrument, which was not called the EQ-5D until 1995, was first presented as a six-dimensional instrument (mobility, self-care, main activity, social relationships, pain and mood). However, following empirical testing, it was very quickly, by 1991, reduced to the five dimensions we recognise today (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). Each dimension of the original EQ-5D, now called the EQ-5D-3L, had three possible response option levels, corresponding to no problems, some problems and extreme problems in each of the five domains. This resulted in $3^5=243$ possible health states (Brooks and The EuroQol, 1996). The measure also included the EQ-VAS, a visual analogue scale (VAS) on which respondents could rate their own global health. Since the aim of the measure was not only to describe health but to value it, an accompanying value set was derived using TTO exercises in 1995 (Dolan, 1997). This resulted in the ability to generate utility values from individuals' responses to the five EQ-5D domains. These utility values range from 1 representing full health, to -0.594, with 0 representing a state equivalent to death.

The EQ-5D has been widely translated with many countries developing their own versions and some also deriving their own country specific tariffs. Despite the popularity of the EQ-5D-3L, questions were raised regarding whether the three levels were sufficient in the measurement of HRQoL (Devlin and Brooks, 2017). Although the validity and reliability of the EQ-5D-3L descriptive system has been demonstrated in many populations and conditions, a substantial ceiling effect has been found especially in general population studies, with many respondents returning maximum scores. In an attempt to reduce this ceiling effect, a five-level version, the EQ-5D-5L was developed (Herdman, Gudex et al., 2011), with severity labels amended to “no problems” “slight problems” “moderate problems”, “severe problems” and “extreme problems”/“unable”. The most extreme level of the mobility item, which had previously been labelled “confined to bed” was replaced by “unable to walk about”. A cross-walk tariff was made available to generate utility values for responses to the EQ-5D-5L based on the EQ-5D-3L tariff (van Hout, Janssen et al., 2012). An EQ-5D-5L value set for England was published in 2017 based on a combination of TTO and DCE methods (Devlin, Shah et al., 2017). The more recent EQ-5D-5L is increasingly being used (Herdman, Gudex et al., 2011) and therefore this analysis will focus on the EQ-5D-5L.

During development of the EQ-5D-5L, the face validity of the new severity labels was tested in the UK and Spain using a VAS response scaling exercise to test respondents' perceptions of the severity labels and focus groups to check the face validity of generated health states (Herdman, Gudex et al., 2011). In the UK the response scaling exercise was carried out in a convenience sample of 40 participants, of which 21 were reported to be over 40 years old and eight were retired/ pensioners. Results suggest that the chosen 5L levels are well distributed across the health continuum and similarly understood in both English and Spanish. For the qualitative testing of face validity, a mix of healthy and patient participants were recruited in the UK. The mean age was 42.5 for the 15 healthy and 43.1 for the 15 patient participants, with one and three reported as retired respectively. We cannot be sure from the reporting how many older adults were involved in this testing, but mean ages around 40, in sample sizes of 30 and low numbers of retired individuals, suggest not many older adults were included. The focus groups had few problems understanding the extent of problems described by the new levels. Face validity was generally clear to participants. It was stated that more work was required to investigate the validity and reliability of the new version. It was also acknowledged that samples were small and had been chosen for convenience and were not representative of the general

population (Herdman, Gudex et al., 2011). Therefore, the generalisability of these results may require checking.

The measurement properties of the EQ-5D-5L compared to the EQ-5D-3L were then investigated in a multi-country study, across members of the general population and eight patient groups (Janssen, Pickard et al., 2013). The total sample size across the six countries was 3,919, with a mean (SD) age of 51.9 (20). The UK sample totalled 1001, with mean ages by condition ranging from 34.3 (ADHD, n=69) to 60.8 (COPD, n=125). There was no further reporting of sample size by age group. This study included the examination of acceptability, convergent and known-group validity and investigation of floor/ceiling effects. Acceptability was examined using completion rates, while ceiling effects were investigated using percentages of respondents returning no problems in all five dimensions (a score of 11111). Known group validity was tested in regard to age, education and smoking status. A lower reported health status was hypothesised with increasing age, lower education levels and in smokers and ex-smokers. Convergent validity was assessed by comparing the dimensions of the two versions of the EQ-5D to the WHO-5 wellbeing questionnaire using Spearman rank order correlations. The ceiling effect was reduced from 20.2% on the 3L to 16.0% on the 5L. Convergent validity with the WHO-5 and known-group validity were confirmed, and convergent validity improved slightly with the 5L (Janssen, Pickard et al., 2013).

3.4.2 ASCOT

The ASCOT was developed as a measure of SCRQoL for the evaluation of social care services. The development of the ASCOT included review of the literature, adaptation of previous measures of outcomes in social care, expert opinion and the direct involvement of social care service users who were involved in providing general advice, as well as cognitive and psychometric testing (Netten, Burge et al., 2012). The ASCOT is originally based on the Older People's Utility Scale (OPUS), a measure of the social care outcomes of older people. This measure was extended using expert opinion and the literature to make it applicable to those aged under 65 and cognitively tested in under 65s. Further changes were then made at several time points, which included dropping several domains which were found not to perform well and the addition of the dignity domain. The format, wording and number of response options

were also altered to reflect capabilities at high levels of functioning and to attempt to reduce large ceiling effects and increase sensitivity at high levels of SCRQoL by having both a “no needs - adequate” and an “ideal” state at the top end of the SCRQoL scale. The wording of domains was also altered to improve their applicability to older adults.

Following these changes, the content validity of the revised instrument, now called the ASCOT, was tested using cognitive interviews with social care service users. Thirty cognitive interviews were carried out in three waves of 10 to allow discussion of any identified issues with items and alteration of wording before the next wave. Approximately half of these respondents were over the age of 65. The wording of several questions and response options were altered and tested in further waves. Although several issues were found in item wording, in general, concepts and response options were reported to be understood as intended (Netten, Burge et al., 2012) and the measure was considered to reflect what was important to participants in relation to their SCRQoL.

A field test of this version of the ASCOT was then carried out to examine further elements of its psychometric performance. The majority of this psychometric testing was based on a sample 301 older (aged 65+) users of home care services taken from the local councils annual User Experience Survey (UES) in 2009. Data was collected through face-to-face computer assisted interviews. Response distributions were examined in order to investigate whether alterations to number and wording of item response options had improved the distribution of responses and sensitivity at the top end of the SCRQoL scale. The distribution of responses was improved. For all items except occupation the distribution was still skewed towards the top end of SCRQoL, however this is to be expected if services are performing well. For five of the eight items (personal cleanliness/comfort, food/drink, accommodation cleanliness/comfort, personal safety and dignity), more than 40% of respondents still responded at the ceiling of the item, which still represents a substantial ceiling and may limit the ability of the measure to discriminate between those with high levels of SCRQoL. The “no needs” levels of the food/drink and accommodation cleanliness/comfort items were further refined to attempt to further reduce ceilings.

Convergent validity was examined between the ASCOT and the EQ-5D measure of health, the 12-item General Health Questionnaire (GHQ-12), a screening measure for psychiatric disorders (Goldberg and Williams, 1988), the control and autonomy

subscale of the CASP-12 (Wiggins, Netuveli et al., 2008) and the ADL and IADL measures of activities of daily living. Moderate correlations were expected between the ASCOT and all these measures, except the GHQ-12, for which a strong relationship was hypothesised. Strong relationships were found between scores on the ASCOT and the GHQ-12, measuring wellbeing and control and autonomy measured with the subscale of the CASP-12 and a moderate relationship was found between the EQ-5D and ASCOT as expected.

An early version of ASCOT, which did not include the dignity item and only had 3 response options per question, was used in two studies, which provided an opportunity for psychometric analysis of item wording and domain choices. The first study included 2,228 individuals from the user experience survey (UES) with physical and sensory impairments, aged 18-64 with a mean age of approximately 50. The second study included 959 individuals receiving individual budgets (IB) for social care, of which n=263 were older people aged over 65. Item and measure level rates of missingness were investigated. In the 263 older people from the IB dataset, at least one item was missing for 15% of the sample, with item level missingness highest for social participation (8.42%) and occupation (5.49%). This is a substantial amount of missing data, particularly at the scale level. High item ceiling effects were found in these analyses. Only 3 items (accommodation, social participation and occupation) had fewer than 50% of respondents returning the top level in the UES sample. Separate response frequencies were not provided for older people in the IB sample, however similar ceiling rates were seen across the whole sample. While ceiling effects could be expected if social care services are being provided well these high levels mean that the measure will be unable to distinguish between respondents with high levels of SCRQoL.

Unidimensionality was also investigated using principal axis factoring (Netten, Burge et al., 2012) on an earlier version of the ASCOT which did not yet include the dignity item and still featured only 3 response levels. Dimensionality was assessed in two samples. The first included 2,228 individuals from the 2007 UES which included individuals aged 18-64 with physical and sensory impairments. The second dataset included 959 individuals receiving individual budgets (IB) for social care, of which n=263 were older people over 65. One factor solutions were obtained in both datasets, with the eigenvalues on the second factor substantially below the Kaiser inclusion rule of one (Kaiser, 1960). In the UES dataset, all factor loadings were strong, with the lowest being 0.625 for occupation. However, one item, occupation,

had high unique variance (>0.6), suggesting the single factor model did not explain this item very well. These results were said to suggest that the items form a weak unidimensional scale. In the IB dataset, factor analysis was conducted on all ages together, with no separate analysis for older adults. Therefore, nothing can be said about the unidimensionality of responses from the perspective of older adults specifically. However, the model did not fit as well as in the UES sample unidimensionality test. All factor loadings were above 0.4, but 3 of the 7 domains had lower loadings than the lowest found in the previous dataset. The control, safety, accommodation and food items all also had unique variances >0.6 suggesting issues in the one-factor model. Again, it was concluded that the items form a weak scale, however the model did not fit as well. However, the current version of the ASCOT has changed quite substantially from this version and therefore this may no longer reflect the dimensionality of the updated ASCOT.

3.4.3 WEMWBS / SWEMWBS

The WEMWBS was developed by a panel of experts from an existing scale, the Affectometer 2 (Kammann and Flett, 1983), which was validated in a sample of the UK general population. Focus groups and psychometric testing of the Affectometer 2 were carried out in order to assess its performance and identify areas for change (Tennant, Hiller et al., 2007). Nine focus groups were carried out, of which one contained exclusively mental health service users and two contained only those aged 65 and above. Participants were asked to complete the Affectometer 2, discuss what positive mental health meant to them and how it related to items on the scale. Content analysis was used to identify items and concepts which participants often found difficult to understand or felt were confusing. Factor loadings and completion rates for each item from an existing survey in the general population were also examined. This evidence, together with reference to the academic literature and their expert opinion helped the development team identify which items should be kept, which reworded, and which dropped or added. This resulted in the 14 item WEMWBS measure, with each equally weighted item presented on a 5-point Likert scale and scores ranging from a minimum level of mental wellbeing of 14 to a maximum of 70.

The WEMWBS was validated in both student and Scottish general population (aged 16+) samples (Tennant, Hiller et al., 2007). In the general population sample there

were 1749 complete responders, of which 274 were aged 65-74 and 61 respondents aged 75+. Scottish general population responses were gathered in face-to-face interview whereas students were given packs to either complete on the spot or to take home and return by post. Acceptability and discrimination were investigated using item-level frequencies of complete responses and response distributions. Structural validity was checked using confirmatory factor analysis to test that the scale was measuring a single underlying concept. Internal consistency was examined by item-total score correlations and using Cronbach's alpha to measure the homogeneity of the global score and assess item redundancy. Test-retest reliability was checked using one-week intra-class correlation coefficients. Floor and ceiling effects were investigated. Convergent and known group validity were examined through correlations between WEMWBS and other measures and testing whether WEMWBS discriminated between known groups in pre-hypothesised ways. In the student sample, of the 354 who responded 98% fully completed the WEMWBS. In the general population sample of 2,075, 16% failed to answer any WEMWBS questions, with non-responders statistically significantly more likely to be older. There were no apparent floor or ceiling effects in either sample. Confirmatory factor analysis supported a single factor model. Cronbach's alpha was 0.89 in the student sample and 0.91 in the general population sample, suggesting some potential item redundancy in the scale (Tennant, Hiller et al., 2007). Hypotheses regarding the relationships between WEMWBS and comparator measures and known-groups were generally confirmed. Test-retest at one week was 0.83 suggesting high reliability.

The face validity of the WEMWBS was checked in two focus groups. These focus groups contained a total of seven participants, all aged 18-64. Participants generally agreed that the WEMWBS was easy to understand and complete and did not suggest any further improvements. However, these focus groups did not include any adults aged 65+ meaning the face validity of the measure itself was not checked in older adults during measure development. The psychometric performance of the WEMWBS has gone on to be checked in several specific groups including young people (aged 13-15) (Clarke, Friede et al., 2011) and Chinese and Pakistani minority populations (Taggart, Friede et al., 2013) in the UK. Both assessed content validity, convergent and structural validity and internal consistency with results largely supporting the psychometric performance of the WEMWBS in these groups.

An investigation into the structural validity of the WEMWBS using Rasch analysis on the Scottish Health Education Population Survey led to the deletion of 7 of the 14

WEMWBS items, which left the 7-item SWEMWBS version (Stewart-Brown, Tennant et al., 2009). Several items were excluded because they did not fit the Rasch model and several due to DiF. It was noted in this study that one of the retained items, “I’ve been feeling optimistic about the future” showed bias for age. This study did not include a qualitative aspect and it is not clear how many respondents to the survey were elderly. Therefore, it is important to further investigate any DiF and qualitative issues with this scale in an older sample.

3.4.4 ONS-4

The ONS-4 questions were developed with the aim of measuring personal subjective wellbeing, as an important component of national wellbeing (Tinkler and Hicks, 2011). In developing the ONS-4, the ONS looked at existing questions in UK and foreign surveys and sought advice from academics, the National Statisticians Advisory Forum and The Technical Advisory Group. The four questions chosen are based on recommendations from Dolan et al (Dolan, Layard et al., 2011). They were chosen to cover a breadth of common types of wellbeing measures; evaluative, experience and eudemonic in an attempt to create a balanced approach to wellbeing measurement. Evaluative appraisals ask individuals to make a cognitive appraisal of their satisfaction of their life as a whole or certain aspects of it (Tinkler and Hicks, 2011). This approach to wellbeing measurement is covered by the question “overall, how satisfied are you with your life nowadays?”. The experience, or affect, approach aims to assess an individual’s emotional quality at a given moment in time (Tinkler and Hicks, 2011). Both positive and negative elements of experience were included with questions “overall, how happy did you feel yesterday?” and “overall, how anxious did you feel yesterday?”. The eudemonic approach includes elements of wellbeing not necessarily captured in the first two approaches such as people’s psychological need to feel their life has meaning or purpose, that they are connected to people and that they have autonomy over their lives (Tinkler and Hicks, 2011). This approach was covered with the question “overall, to what extent do you feel the things you do in your life are worthwhile?”. These questions were intended to be used for overall monitoring rather than specific policy appraisal.

Content validation work was undertaken in a sample of 44 participants (Ralph, Palmer et al., 2011). Eight questions were discussed, the ONS-4 and four more questions being considered for inclusion following previous user feedback. Purposive sampling

was used with primary stratifiers of sex, age and socioeconomic group. A combination of face-to-face in-depth and cognitive interviews were used in three waves. Of these 44 participants, eight were aged 61 or over. Reaction to and understanding of the ONS-4 questions was mixed and issues with questions were identified. For example, for the question "overall, how satisfied are you with your life nowadays?" the term "satisfied" was not uniformly understood and was sometimes seen as a negative or neutral state, where something was neither good or bad and was therefore considered not something to aim for. In the third and final wave of interviews "content" was tested as an alternative to "satisfied". This term was considered comparable to satisfied but less likely to be seen negatively. This suggested "content" to be a viable alternative if further problems were found with satisfied. The term "nowadays" was found old fashioned and either not understood or ignored. Further issues were identified with the other questions including the most vulnerable respondents becoming visibly upset when answering "overall, to what extent do you feel the things you do in your life are worthwhile?" (Ralph, Palmer et al., 2011). Confusion over the layout of the anxiety question was also seen. As this is the only negatively phrased item, the direction in which respondents signal higher levels of wellbeing reverses, however some failed to notice and provided inconsistent answers. Despite these problems the ONS-4 questions have not been altered and have advanced essentially unchanged.

Wording variations, question order effects and factors associated with the ONS-4 questions were tested in the OPN surveys (Office for National Statistics, 2012). Item non-response was described as low. No evidence was found of psychometric testing of the ONS-4 questions.

While these four questions are grouped together in the personal wellbeing section of the overall ONS wellbeing programme, there is no guidance to suggest they should be classed as a single measure or summed. Despite this there is a growing tendency to do so and therefore these four questions will be considered as a single measure of wellbeing in order to investigate whether this is statistically and qualitatively appropriate in an older population.

Little work has been done, either statistically or qualitatively, to check the validity of the ONS-4 questions, especially in subgroups such as the elderly. In the qualitative work described above only eight respondents were 61 and over (Ralph, Palmer et al., 2011). Therefore, it was considered important to include this measure as this is considered an important gap in current research. There have also been issues with

patient reactions in a local collection of ONS-4 responses during face-to-face interviews with frail elderly participants in a Vanguard project in Wakefield (Healthwatch Wakefield, 2016). This has emphasised the need to establish the validity and appropriateness of the ONS-4 in an older population.

3.4.5 SF-12v2

The SF-12 is a reduced 12-item version of the SF-36. The SF-36 was developed in 1990 with the aim of creating a short form measure of health which was feasible and acceptable to respondents in terms of burden, but which provided a detailed and comprehensive coverage of what is important to health (Maruish, 2012). The SF-36 covered eight domains; general health, physical functioning, physical role functioning, emotional role functioning, bodily pain, vitality, mental health and social functioning. Each domain contained between two and ten items, with various numbers of response options (between two and six). Domain scores could be summed to generate physical and mental component scores (PCS and MCS) (Ware, Snow et al., 1993). The PCS covered general health, physical functioning, physical role functioning and bodily pain while the MCS covered vitality, mental health and social functioning. Items for the SF-36 were based on existing measures of health available in the literature and judgement of the SF-36 developers (Maruish, 2012). It is not clear whether any patients or members of the public were consulted during the generation of domains and items. Preliminary versions of the SF-36 were tested in large field surveys for two years which allowed for testing of psychometric performance. In these field tests respondents were also able to suggest improvements (Maruish, 2012) and therefore there was patient/public input into instrument refinement, although the coverage of domains were not added to or changed as a result of this. Changes, outlined below, were to simplify layout and wording and alter response formats.

In 1994 the SF-12 was constructed with the aim of creating a shorter version of the SF-36 which could reproduce its MCS and PCS scores using only a subset of the items, with each of the eight domains represented by only one or two items and response scales which matched the SF-36 format (Maruish, 2012). Regression methods were used to select the subset of items from the SF-36 and create weighting algorithms to reproduce the SF-36 PCS and MCS scores (Maruish, 2012).

Following publication of the SF-36 it became clear from data, respondent feedback during field tests and qualitative research that improvements could be made. Qualitative research was said to include "numerous focus group studies and formal cognitive tests" (Maruish, 2012), however no further detail on the methods, sample or even countries in which this research took place were provided. These improvements led to both the SF-36v2 and SF-12v2. Changes included revision and shortening of the wording of instructions and questions to simplify and improve understanding (Maruish, 2012). The layout was also amended to ease completion and reduce missing responses. The dichotomous response options used for the items in the physical and emotional role domains were changed to five response option Likert scales, which would substantially reduce floor and ceiling effects seen within existing data for these items and increase their range, while the six response options of the mental health and vitality items were amended to five options in response to evidence of confusion about the ordering of and difference between some of these response options (Maruish, 2012). Scoring methods were also amended. All eight health domains were incorporated into the score of both the PCS and MCS based on IRT methods, which were used to identify their factor loadings for these two constructs (Maruish, 2012). Norm-based scoring was also introduced for both the domain scores and the PCS and MCS component scores, using t-score transformation methods based on US general population norms (Maruish, 2012). Following these t-score transformations the domain and component scores each have a mean of zero and a standard deviation of 10 (Maruish, 2012).

Although psychometric testing and adaptations to the SF-36 and SF-12 were made at several points, using both qualitative and quantitative methods, it is not clear how many older adults were included in either type of testing. It is also not clear whether at any time the development team investigated whether the SF-12 or SF-36 formed a comprehensive assessment of what individuals of any age felt was important to their health using qualitative methods or whether they checked the relevance or acceptability of included items.

From details provided by development teams it does not appear that older adults were involved in the development of the content of the EQ-5D, SF-12v2 or ONS-4. While older adults were consulted in the early stages of the development of the WEMWBS, they were only asked to complete and discuss the Affectometer 2, the source measure on which the development of the WEMWBS was based. Substantial changes were made between this measure and the resulting WEMWBS, but the content validity of

the WEMWBS was never confirmed in older adults. The content validity of the ONS-4 was checked in a range of age groups after its release, of which eight of 44 participants were aged 61+. Issues were identified, particularly in those participants described as most vulnerable, however age details were not provided. Despite the issues identified, the ONS-4 questions proceeded unchanged. The content validity of the ASCOT was checked in older adults at several stages during its development, making this the measure in which we can be most confident of content validity in older adults. While statistical psychometric field testing of the included measures tended to include at least a small proportion of older adults, results were rarely provided according to age subgroups, meaning we cannot be sure of the psychometric performance in this age group. Testing of the structural validity of the WEMWBS identified age related DiF within the measure, but failed to provide further details, leading to questions regarding the performance of this measure across age groups. The exception to this is again the ASCOT, which provided separate psychometric testing and results in older adults.

This section provided details of each measure and the process by which they were developed and tested. Any input from older adults during the process was noted as this is an important starting point from which to ascertain the evidence of the psychometric performance of these measures in older adults. The next section will examine the existing literature to investigate the evidence of their performance in older adults, an important group who use a disproportionately large amount of health and social care resources, which will continue to grow in the context of population ageing, and who are often underrepresented in outcome research.

3.5 – Systematic review investigating the psychometric performance of the EQ-5D, SF-12, ASCOT, WEMWBS and ONS-4 in older adults

3.5.1 Introduction

This section presents a systematic review of the existing evidence on the psychometric performance of the chosen PROMs in measuring the health, QoL and wellbeing of older adults. First the methods chosen to identify relevant studies, assess the quality of those included studies and synthesis this evidence are outlined as well as the methods and criteria for classifying whether or not an instrument performs well in each psychometric property. Then the results are presented and discussed in terms of the key findings and limitations.

3.5.2 Systematic review question

This systematic review will investigate the existing evidence on the psychometric measurement properties of the EQ-5D, ASCOT, ONS-4, SF-12 and WEMWBS (or SWEMWBS) in assessing the health, QoL and wellbeing of older people.

3.5.3 Methods

3.5.3.1 Search strategy

The electronic search strategy is outlined in Appendix 3. It combines terms for relevant psychometric properties such as validity, reliability and acceptability; the main theories for testing measurement performance; a variety of terms for elderly populations and all identified name combinations for the selected QoL measures. This search strategy was used in a variety of online databases including PubMed, MEDLINE, CINAHL and the Cochrane Library as they were included in similar reviews of the psychometric performance of PROMs in healthcare evaluation (Bulamu, Kaambwa et al., 2015, Haywood, Brett et al., 2017) and thought to comprehensively span the relevant health economics outcomes literature. These databases were searched from their inception to September 2017. Reference lists of included papers were hand searched for further relevant papers.

3.5.3.2 Inclusion and exclusion criteria

The inclusion and exclusion criteria were as follows

Inclusion criteria

- Validation must be completed for either EQ-5D-3L, EQ-5D-5L, ASCOT, ONS-4, SF-12, WEMWBS or SWEMWBS
- Studies investigating the psychometric performance of the descriptive system of the relevant measures in terms of any of the following properties; validity, reliability, responsiveness or acceptability
- Validation population must be aged 60 years and over.

Exclusion Criteria

- Studies where the measure is only applied in older people with no psychometric assessment of measurement performance
- EQ-5D-3L, EQ-5D-5L, ASCOT, ONS-4, SF-12, WEMWBS or SWEMWBS is solely used as a comparison measure in the validation of another measure and is not itself being validated
- Studies not available online
- Studies not available in English
- Validation of valuation technique rather than a validation of a descriptive system

All titles and abstracts were screened for inclusion by the researcher. A random 10% of title and abstracts returned by the electronic search were double reviewed by a supervisor (CH) and agreement was checked. There were no exclusions based on study design. Any published work investigating the psychometric performance of these measures, quantitatively and/or qualitatively, in older adults (aged 60 years and above) was included. Studies which only applied the measure in an older population, without assessing the psychometric properties of the measure were excluded. Where reviews were found, relevant included papers were screened at full text. While this thesis focuses on the UK perspective, international validation in older adults was included, although not considered fully generalizable to a UK population. This includes work validating translated versions of the relevant measures.

Validations in specific condition groups, for example fracture patients, were included as long as the sample were over 60 years old. Several returned papers were looking

at the validity of the use of these measures in older people with dementia. These often focussed on the validity of proxy responses to the included measures in terms of their agreement with patient reported scores. If there was no additional validation of the measure itself beyond the investigation of patient, proxy agreement these papers were not included as this was not considered sufficient information of the measurement performance of the measures themselves in older adults. However, if additional aspects of the validity, reliability or responsiveness of these measures in dementia patients were examined these were included.

3.5.3.3 Data extraction and quality appraisal

Data was extracted on study design, including study population, setting, mode of administration of the measure and evidence found for measurement properties including: validity (structural, construct (convergent and known group), content and cross-cultural), reliability (internal consistency, test-retest and inter-rater), responsiveness and interpretability (rates of missing data and how missing data was handled, response distributions including ceilings and floors, minimally important differences and mean and change scores). Data extraction followed the structure of a recent review of the psychometric properties of PROMs in hip fracture patients (Haywood, Brett et al., 2017), and the COSMIN taxonomy and critical appraisal checklist used in this study (Mokkink, Terwee et al., 2010, Terwee, Mokkink et al., 2012).

The COSMIN checklist for systematic reviews of measurement properties, shown in Appendix 4, was used to assess the methodological quality of included studies (Terwee, Mokkink et al., 2012). The checklist contains measurement property specific checklists each of which contains a list of items against which the methodological quality of included studies can be assessed on a 4-point scale (excellent, good, fair, poor). The methodological quality of each included study, in relation to each measurement property assessed, is determined by the lowest ranking given to any item in the relevant measurement property specific box in the checklist.

3.5.3.4 Assessment of psychometric results

Construct validity

Two commonly assessed elements of construct validity are convergent validity and known group validity (Fayers and Machin, 2016). Convergent validity assesses the extent to which measures which aim to measure the same or similar underlying constructs or concepts agree with each other. This is commonly measured using correlations, either between overall measure scores or utilities, or dimension or item scores. The strength of correlations is used to judge the extent to which measures are related to each other. Different studies use different cut-off systems to label the strength of correlations. This review used Cohen's d criteria which labels correlations as either trivial (correlations < 0.2), small ($0.2 \leq$ correlations < 0.5), moderate ($0.5 \leq$ correlations < 0.8) or large (correlations ≥ 0.8) (Cohen, 1988). The expected direction and magnitude of correlations between measures should be hypothesised a-priori according to the COSMIN criteria (Terwee, Mokkink et al., 2012).

Known group validity assesses the extent to which a measure can distinguish between groups hypothesised to differ in the underlying construct or trait (Fayers and Machin, 2016), such as patients vs general population or individuals with different severities of a condition. It can be hypothesised that those with poorer health will have worse scores on a measure. These hypotheses can then be tested using appropriate statistical tests to examine whether there is a statistically significant difference in the scores of these different known groups in the expected direction and magnitude.

Another element of construct validity is measurement invariance, which requires that items perform the same in different subgroups of people. Where the probability of responding in a certain category varies accord to characteristics other than the respondent's level of underlying trait, measurement invariance is violated, and the item is said to exhibit DiF. DiF can be examined using ordinal regression methods, multiple group IRT models or structural equation models.

Structural validity

Structural validity is defined as the degree to which the scores of a measure are an adequate reflection of the dimensionality of the construct to be measured (Mokkink, Terwee et al., 2010). Structural validity is commonly assessed using factor analysis. EFA can be used to explore the number of potentially meaningful factors, representing distinct underlying concepts, within a measure while CFA can be used to examine the extent to which the data fit the predefined measurement model of a measure and

explore whether there are alternative, more suitable factor structures which fit the data better. The intended factor structure of the measure should be clearly set out by the measure developers. EFA can be run to see if this structure is suggested to best fit the data. Then CFA should be run to assess the model goodness of fit of this structure to the data. Common cut-offs for good CFA model fit are CFI or TLI >0.95, Root mean square error of approximation (RMSEA) <0.06 or Standardised root mean residuals (SMRM) <0.08 (Terwee, Bot et al., 2007). Model fit can also be compared between alternative CFA model structures and the intended model factor structure. Similarly to CFA methods, structural validity in terms of factor structure and model fit can also be assessed using IRT models.

Content validity

Content validity is the degree to which the content of a PROM is an adequate reflection of the construct to be measured. This is assessed using qualitative methods in members of the population in which the measure will be used. This qualitative work should investigate whether all items are relevant to the target population and to the construct being measured and whether the measure comprehensively covers the important aspects of that construct (Terwee, Bot et al., 2007).

Internal consistency

Internal consistency is the degree of interrelatedness among items (Mokkink, Terwee et al., 2010). Internal consistency is commonly measured in CTT using Cronbach's alpha. The common cut-off rule for good internal consistency of a scale is that the alpha should be $0.70 \leq \alpha \leq 0.95$ (Terwee, Bot et al., 2007), as values below 0.7 may suggest that the items are not sufficiently related to be measuring a single concept, while $\alpha \geq 0.95$ may suggest item redundancy, with items excessively similar and in fact asking the same thing. Internal consistency should be measured separately for each factor or dimension within a multidimensional measure. Internal consistency can also be assessed using IRT methods by examining the level of information provided by items and the measure as a whole and how this varies across individuals with different levels of underlying trait. Total measure information can be used to assess internal consistency reliability as total measure information is analogous to estimating Cronbach's alpha at each point on the QoL scale, with Cronbach's alpha=1-(1/total information) (Petrillo, Cano et al., 2015).

Test retest and Inter-rater reliability

Reliability is defined as the degree to which the measurement is free from error. Test retest reliability examines the extent to which scores for patients who have not changed remain the same over repeated measurements at different times. Inter-rater reliability examines the extent to which the scores different people assign to the health of an individual at the same point in time agree. These elements of reliability can be examined using intraclass correlation coefficients (ICCs), weighted kappas, with cut-offs for good levels of these being ≥ 0.7 (Terwee, Bot et al., 2007).

Responsiveness

Responsiveness is defined by the COSMIN group as the ability of a PROM to detect change over time in the construct to be measured (Mokkink, Terwee et al., 2010). In the literature, responsiveness is commonly assessed by examining whether there is a statistically significant difference in scores obtained before and after a change is expected to have occurred, for example before and after a successful treatment. The magnitude of the change in scores can be assessed using standardised effect sizes (SES) or standardised response means ($SRM = \text{the change in score} / \text{the change standard deviation}$). Again, the expected direction and magnitude of change should be hypothesised a-priori.

However, the COSMIN taxonomy has a preference for other methods of judging responsiveness (Mokkink, Terwee et al., 2010). These are criterion-based and construct-based assessments of responsiveness. Criterion based methods assess whether there is a statistically significant difference in the change scores of groups over time who do and do not meet the predefined external criteria, given the a-priori hypothesis that these groups will have change scores of different directions or magnitudes. Construct based assessments of responsiveness compare the change scores between measures aiming to measure the same or similar constructs. Change scores can be compared by comparing mean differences of change scores or examining correlations between the change scores. Again, hypotheses about the expected direction and magnitude of correlations or mean differences of change scores should be made a-priori.

3.5.3.5 Data synthesis

Data relating to each measurement property for each included PROM was synthesised. Data synthesis accounted for several factors including: the number of studies reporting evidence on that measurement property for that PROM; the methodological quality of those studies, as judged by the COSMIN scores; the results for each measurement property and the consistency of results between studies. In line with previously published reviews of measurement properties of QoL measures (Elbers, Rietberg et al., 2012, Haywood, Brett et al., 2017, Haywood, Collin et al., 2014) which also used the COSMIN checklist, each measurement property for each PROM was given a score made up of two parts. First was a rating of the overall quality of the measurement property based on the results seen. This score, referred to as “results”, could be given as: + “adequate”, - “not adequate”, +- “conflicting or ?” “unclear”. The second part of the score, referred to as “thoroughness”, was the level of evidence on which the overall quality of that measurement property was based. This could be rated as: “strong” if consistent findings were seen across multiple studies of good methodological quality, or in one excellent quality study; “moderate” if consistent findings were seen in multiple studies of fair methodological quality, or one good quality study; “limited” if based on one study of fair methodological quality; “conflicting” if finding were conflicting and “unknown” if evidence was based solely on studies of poor methodological quality.

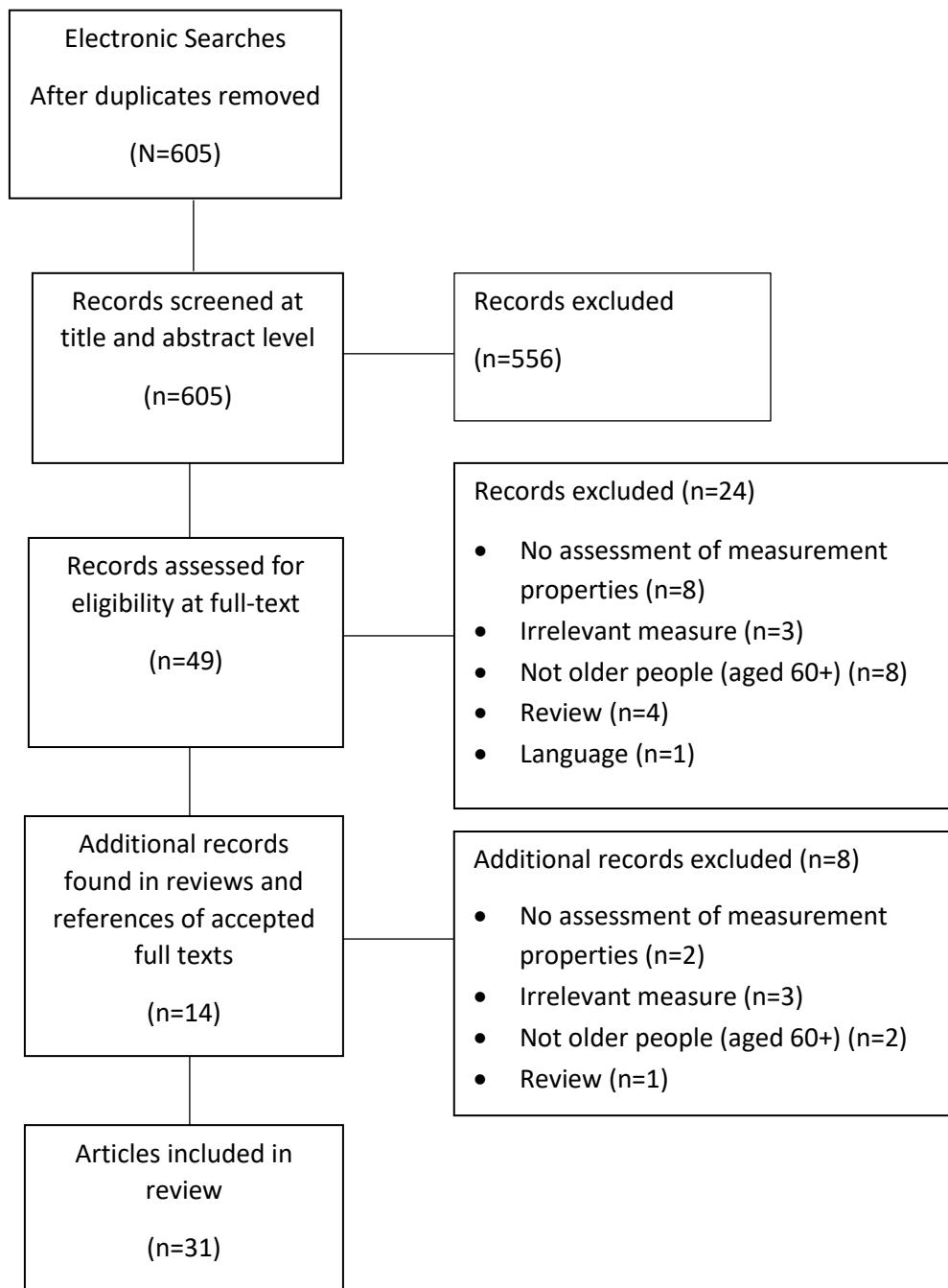
3.5.4 Results

3.5.4.1 Identified studies

The PRISMA flow diagram in Figure 5 provides a visual representation of the number of articles obtained, included and excluded at each stage of the search. The electronic search strategy, run on September 18th 2017, returned 605 articles after duplicates were removed. After title and abstract screening 49 papers were screened at full text level. Of these 25 were included in the review. An additional 14 papers were identified through hand searching the reference lists of included papers, of which 6 were included. A list of included studies can be found in Appendix 5. Papers were excluded for the following reasons: not assessing psychometric measurement properties (n=10), not assessing the measurement properties of one of the included measures or the included measure was only a comparator in the validation of another measure (n=6), the evaluation of psychometric measurement properties was not based on

samples over the age of 60 years (n=10) and study was a review (n=5) and the study was written up in a language other than English (n=1). Thirty-one papers were included in the review in total. There were several instances where multiple papers described the same validation study in the same sample. These were combined leaving 28 validation studies included in the literature review.

Figure 5 – PRISMA Flow Diagram



3.5.4.2 Characteristics of included studies

The characteristics of included studies are summarised in Appendix 6. Most of the included studies investigated one or more of the measurement properties of the EQ-5D-3L (n=17) and the SF-12 (n=10) in older people. Four studies investigated the ASCOT and one the EQ-5D-5L. This is partially to be expected since these measures are much more recently developed than the EQ-5D-3L and SF-12. No papers were found to assess the measurement properties of the WEMWBS or ONS-4 in older people specifically.

Seven studies were set in the UK, four each in the USA and Sweden, three in Australia, two in China, two in the Netherlands and one each in Canada, Spain, Germany, Mexico and Israel. Sample sizes ranged from 25,637 in an analysis of a large public dataset which included the EQ-5D-3L in the Netherlands (Lutomski, Krabbe et al., 2017) to 60 in a study which investigated the responsiveness of the EQ-5D-3L in older women with femoral neck fractures in Sweden (Tidermark and Bergström, 2007) and 10 in a qualitative study of the content validity of the Dutch translations of the EQ-5D-3L and ASCOT in the Netherlands (van Leeuwen, Jansen et al., 2015). Most studies were populated with community dwelling older people, while fewer focussed on elderly living in residential and nursing homes. Some studies investigated measurement properties of PROMs in general samples of older people (n=17), while others recruited older people with a specific condition (n=9). Two studies considered condition specific groups alongside a general population group (Jakobsson, Westergren et al., 2012, Resnick and Parker, 2001).

The vast majority of the included papers based their investigation of measurement properties on CTT methods (n=27) over IRT methods (n=1) and qualitative methods (n=1). Most studies examined construct validity (n=25), in terms of convergent validity (n=17) and known group validity (n=16). Quality ratings for construct validation were mixed as often hypotheses were vague resulting in fair quality appraisal ratings. Internal consistency (n=11) and structural validity (n=10) were also often examined, but the bulk of this evidence was provided for the SF-12 and was much less frequently reported for the EQ-5D and ASCOT. A lack of clear reporting of factor analytic methods used and the resulting model fit as well as broader methodological issues often lowered the quality rating for structural validity while the internal consistency quality rating was often let down by failing to conduct internal consistency in each separate unidimensional factor, following either tests of structural validity or the hypothesised structure of the measure. There was only one assessment of the

content validity of a measure in older adults, performed on the Dutch versions of the EQ-5D-3L and ASCOT, judged to be of excellent quality. There were some assessments of test-retest and interrater reliability (n=6) and responsiveness (n=9) but these were often vague in their reporting and methodological issues meant they were often rated as fair.

3.5.4.3 Quality of included studies

A table summarising the COSMIN ratings for methodological quality given to each study in relation to each measurement property assessed can be found in Appendix 7. The most common issue which resulted from the COSMIN quality appraisal tool was that studies were rarely clear on the handling of missing data. However, it could usually be deduced or assumed how missing items were handled and therefore the good quality rating was usually chosen for this aspect, unless no comment at all was provided about missing data, to avoid the vast majority of studies being rated poor solely due to this aspect, which leaves little room to easily further distinguish the quality of studies.

3.5.4.4 Measure specific findings

The results obtained from each included study regarding the measurement properties of the PROMs assessed are summarised in Appendix 8. The data synthesis of the overall level of evidence found in relation to each measurement property for each PROM and whether this evidence suggests that the PROM is adequate or not in relation to each measurement property or whether this remains unclear, is shown in Table 3 below.

EQ-5D-3L and EQ-5D-5L

The EQ-5D-3L is the most widely validated measure in older adults, with 16 studies examining performance of the EQ-5D-3L in relation to at least one measurement property in older adults. Nine of these were randomised studies. Eight studies examined the performance of the EQ-5D-3L in older adults with a specific condition (dementia=2 (Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Kunz, 2010),

fracture=4 (Parsons, Griffin et al., 2014, Tidermark and Bergström, 2007, Tidermark, Bergström et al., 2003, Tidermark, Zethraeus et al., 2002), frailty=1 (van Leeuwen, Bosmans et al., 2015b) and mobility impairment=1 (Davis, Bryan et al., 2012)) while four were populated with general older adults, two in those with recent hospital stays (Coast, Peters et al., 1998, Holland, Smith et al., 2004), one in recipients of home care services (Kaambwa, Gill et al., 2015) and two studies compared condition specific groups with either general older adults (Sanchez-Arenas, Vargas-Alarcon et al., 2014) or a “healthy” group (Lutomski, Krabbe et al., 2017). Ten studies included only older adults living in the community, two studies included only those living in nursing or residential care and four did not state the living situation of their sample. The minimum age of participants in these studies ranged from 60 years old (Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Liang and Wu, 2014, Parsons, Griffin et al., 2014, Sanchez-Arenas, Vargas-Alarcon et al., 2014) to 80 (Holland, Smith et al., 2004).

Table 3 – Data synthesis

		EQ-5D-5L	EQ-5D-3L	SF-12	ASCOT	(S) WEMWBS	ONS-4
No. evals in older adults	Total	1	17	10	3	0	0
	UK	0	6	1	1	0	0
Content Validity	Thoroughness	None	Moderate	None	Moderate	None	None
	Results		+		+		
Construct validity	Thoroughness	Moderate	Strong	Strong	Strong	None	None
	Results	+	+	+	+		
Structural validity	Thoroughness	None	Limited	Conflicting	Unknown	None	None
	Results		?	+/-	?		
Test-retest Reliability	Thoroughness	None	Limited	Limited	Limited	None	None
	Results		+	?	+		
Inter-rater Reliability	Thoroughness	None	Conflicting	None	None	None	None
	Results		+/-				
Internal consistency	Thoroughness	None	Limited	Strong	None	None	None
	Results		?	+			
Responsive- ness	Thoroughness	None	Moderate	None	Limited	None	None
	Results		+		?		

There was strong evidence of the construct validity of the EQ-5D-3L, which was examined in 12 studies, all of which used CTT methods to assess construct validity in terms of either convergent or known group validity, or both. Nine studies assessed

the convergent validity of the EQ-5D-3L while eight assessed known group validity. Two convergent validity studies were carried out in the Netherlands (Lutomski, Krabbe et al., 2017, van Leeuwen, Bosmans et al., 2015b), two in the UK (Coast, Peters et al., 1998, Parsons, Griffin et al., 2014) and one each in Canada (Davis, Bryan et al., 2012), Australia (Kaambwa, Gill et al., 2015), Spain (Diaz-Redondo, Rodriguez-Blazquez et al., 2014), Mexico (Sanchez-Arenas, Vargas-Alarcon et al., 2014) and Germany (Kunz, 2010). Of the ten studies which investigated convergent validity five studies were conducted in condition specific samples (dementia=2 (Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Kunz, 2010), fracture=1 (Parsons, Griffin et al., 2014), frailty=1 (van Leeuwen, Bosmans et al., 2015b) and mobility impairment=1 (Davis, Bryan et al., 2012)) while one study considered recent hospital patients (Coast, Peters et al., 1998), one care service recipients (Kaambwa, Gill et al., 2015), one members of the general population compared to those with dementia (Sanchez-Arenas, Vargas-Alarcon et al., 2014) and one a variety of conditions and a healthy group (Lutomski, Krabbe et al., 2017). Six studies of convergent validity were rated as good (Coast, Peters et al., 1998, Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Kaambwa, Gill et al., 2015, Kunz, 2010, Lutomski, Krabbe et al., 2017, van Leeuwen, Bosmans et al., 2015b), two as fair (Davis, Bryan et al., 2012, Parsons, Griffin et al., 2014) and one as poor (Sanchez-Arenas, Vargas-Alarcon et al., 2014).

Common comparison measures for convergent validity tests were measures of functional status and activities of daily living (Barthel Index, IADLs, Katz ADLs), other measures of health and QoL (SF-36/SF-12 component scores, ICECAP-O, ASCOT) and measures relating to the specific conditions being assessed by some of the studies (MMSE, QUALID and QOL-AD for dementia, PPA and SPPB for mobility impairment and Oxford hip score and specific questions about pain and mobility for fracture). As would be expected, EQ-5D scores and dimensions were more closely related to other measures of health, activities of daily living and function status and less closely related to broader measures of QoL such as ASCOT and ICECAP-O. Condition specific measure comparisons generally found correlations with the relevant domains of the EQ-5D but weak correlations were found between the EQ-5D and MMSE in studies of dementia patients (Davis, Bryan et al., 2012, Kunz, 2010, Sanchez-Arenas, Vargas-Alarcon et al., 2014).

Of the eight studies which investigated known group validity, three were carried out in the UK (Brazier, Walters et al., 1996, Coast, Peters et al., 1998, Holland, Smith et al., 2004), and one each in Australia (Kaambwa, Gill et al., 2015), Spain (Diaz-

Redondo, Rodriguez-Blazquez et al., 2014), Mexico (Sanchez-Arenas, Vargas-Alarcon et al., 2014), Sweden (Tidermark, Zethraeus et al., 2002) and the Netherlands (Lutomski, Krabbe et al., 2017). One convergent validity study was conducted in a condition specific dementia sample (Diaz-Redondo, Rodriguez-Blazquez et al., 2014) and one in a fracture patient sample (Tidermark, Zethraeus et al., 2002), two studies considered recent hospital patients (Coast, Peters et al., 1998, Holland, Smith et al., 2004), one care service recipients (Kaambwa, Gill et al., 2015), one study used members of the general population (Brazier, Walters et al., 1996), one compared the general population to those with dementia (Sanchez-Arenas, Vargas-Alarcon et al., 2014) and one a variety of conditions and a healthy group (Lutomski, Krabbe et al., 2017). Five studies including known group validity tests were rated as good (Brazier, Walters et al., 1996, Coast, Peters et al., 1998, Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Kaambwa, Gill et al., 2015, Lutomski, Krabbe et al., 2017), two as fair (Holland, Smith et al., 2004, Tidermark, Zethraeus et al., 2002) and one as poor (Sanchez-Arenas, Vargas-Alarcon et al., 2014).

Common characteristics on which tests of known group validity were based were age, education level, living situation, presence of self-reported long-term conditions and comorbidities, measures of recent service use such as GP visits and hospital inpatient/outpatient/A and E attendances, receipt of informal care and whether the EQ-5D could discriminate between those who did and did not have conditions such as dementia and depression based on cut-off scores from other measures. Significant relationships were found for many of the tests relating directly to health status and service use, with EQ-5D scores significantly higher for those with reported higher levels of general health (Kaambwa, Gill et al., 2015) and higher functional status (Diaz-Redondo, Rodriguez-Blazquez et al., 2014) and significantly higher EQ-5D scores in those with recent GP visits and hospital inpatient stays (Brazier, Walters et al., 1996), those with long-term conditions and comorbidities (Brazier, Walters et al., 1996, Coast, Peters et al., 1998, Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Lutomski, Krabbe et al., 2017), those with higher functional status (Diaz-Redondo, Rodriguez-Blazquez et al., 2014) and those with depression (Diaz-Redondo, Rodriguez-Blazquez et al., 2014). However, relationships with demographic variables were much less clear and often not significantly different across groups. The relationship between EQ-5D scores and the age of the older respondent was more complicated. While two studies reported significantly lower EQ-5D scores in older age groups (Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Lutomski, Krabbe et al., 2017), two studies actually saw mean EQ-5D scores rising in older age groups

(Brazier, Walters et al., 1996, Kaambwa, Gill et al., 2015) with this relationship being significant in one of these studies (Kaambwa, Gill et al., 2015).

Limited evidence of the structural validity of the EQ-5D-3L was found. Only one Spanish study, judged to be of good quality (Diaz-Redondo, Rodriguez-Blazquez et al., 2014), was found which assessed the structural validity of the EQ-5D-3L. The study population was made up of institutionalised older adults with dementia, where the EQ-5D-3L was completed by the carer. EFA and PCA were used to examine the dimensionality and unidimensionality was tested using a Rasch model. Two factors were found. The first, a functional factor, included mobility, self-care and usual activities and the second, a subjective factor included pain and anxiety and depression. Lack of unidimensionality was confirmed by lack of fit in the Rasch model. This evidence was not considered sufficient to be able to judge the structural validity of the EQ-5D in older adults and therefore this measurement property was given a rating of unknown in the data synthesis. A potential reason for the limited number of studies investigating the structural validity of the EQ-5D is that it is conceptualised as a multidimensional measure. Therefore evidence of multidimensionality would not invalidate the measure. However it is still important to investigate how the items/dimensions relate to each other, to form a measure of HRQoL, in practice and therefore tests of structural validity and dimensionality are still of interest, even if evidence of multidimensionality would not be a concern.

Moderate evidence was found of the content validity of the Dutch EQ-5D-3L in community dwelling frail older adults was examined in one study based in the Netherlands using think aloud interviews (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Jansen et al., 2015). This study was judged to be of excellent quality and used think aloud interviews in a maximum variance sample of 10 older adults to compare the content validity of the Dutch versions of the EQ-5D-3L, ASCOT and ICECAP-O. The EQ-5D-3L was generally well understood, and often found to be the easiest to answer of the three measures, as the questions were more specific. However usual activities, pain/discomfort, and anxiety/depression were often interpreted more narrowly than intended by the measure developers and the interviewer often perceived positive answering, where the respondent gave answers which suggested higher levels of HRQoL than was expected by the interviewer, given their knowledge of the respondent. However, content validity is known to vary across cultures due to different preferences, attitudes and interpretations of questions, often

due to translation effects. Therefore, we cannot be sure that these results apply to the English version of the EQ-5D.

Limited evidence of internal consistency reliability was found in three studies of fair quality, set in Mexico (Sanchez-Arenas, Vargas-Alarcon et al., 2014), China (Liang and Wu, 2014) and Spain (Diaz-Redondo, Rodriguez-Blazquez et al., 2014). The Spanish study sample was comprised of institutionalised older adults with dementia, the Chinese study sampled community dwelling older adults and the Mexican study sampled general older adults of which 5% had dementia and results were presented separately. A common issue in these studies was the dimensionality of the EQ-5D and how to calculate Cronbach's alpha accordingly. According to the COSMIN checklist Cronbach's alpha dimensionality should be tested, or a test should be referred to from another study in a similar population (Terwee, Mokkink et al., 2012). Cronbach's alpha should then be calculated for each unidimensional subscale. The Spanish study calculated Cronbach's alpha for the EQ-5D-3L as a whole despite having found two factors in unidimensionality tests but item total correlations per item were provided. The Chinese study assumed the EQ-5D to be five dimensional according to the description of the measure and calculated Cronbach's alpha per item and for the measure as a whole. However, in the structural equation modelling section of the paper the EQ-5D was treated as unidimensional and it is not clear whether alternative factor structures were tested. In the Mexican paper, again the EQ-5D was assumed 5 dimensional according to the measure description but this was not tested. Separate Cronbach's alpha statistics were provided for each item. Values of Cronbach's alpha provided were over 0.7, the common cut-off for good internal consistency, except in the Spanish dementia sample where the scale Cronbach alpha=0.64. Again, the multidimensional structure may cause questions as to the relevance of tests of internal consistency for the multidimensional EQ-5D. Similar to structural validity, it can be argued that there is still value in understanding how items relate to each other and to the overall concept of HRQoL which they seek to measure.

Test retest and inter-rater reliability were assessed by two studies each. There was limited evidence of test-retest reliability, assessed in a UK study of women aged 75+ recruited from an RCT of clodronate, a drug aiming to reduce the incidence of hip fracture (Brazier, Walters et al., 1996), which was judged to be of fair quality, and a Dutch study of community dwelling frail older adults, judged to be of good quality (van Leeuwen, Bosmans et al., 2015b). The English study reported correlations between administrations in those who stated that their health had not changed in the 3-month

period and mean and variances of the differences in scores. A fair quality rating was given to this study as a 3-month gap between administrations was considered potentially inappropriate to investigate test retest reliability as over longer periods we can be less sure that no change in health occurred. The Dutch study used a gap of 1 – 2 weeks, over which it is much more likely that no significant change occurred. The Dutch study reported ICCs and weighted Kappas between administrations. The English study found strong correlations and insignificant difference in utility scores while the ICCs and Kappas passed acceptable cut-offs in the Dutch study.

There was conflicting evidence of inter-rater reliability, which was examined in a Spanish study of institutionalised dementia patients (van Leeuwen, Malley et al., 2014) and a German study of community dwelling dementia patients (Kunz, 2010), both judged to be of fair quality. No information was provided about the test conditions or administration for the Spanish study to enable readers to feel confident that administrations were similar while in the German study one group answered face-to-face and the other through telephone interviews which meant administrations were not similar. Results for inter-rater reliability were presented using ICCs in both papers but results were mixed, with the Spanish paper passing the 0.7 cut-off for acceptable inter-rater reliability but the ICC from the German paper being substantially lower. Weighted kappas for agreement between raters for each item ranged from mild to moderate in the German study.

Moderate evidence of responsiveness of the EQ-5D-3L in older adults was found. Responsiveness was measured in eight studies, of which four were judged to be of good quality (Kunz, 2010, Parsons, Griffin et al., 2014, Tidermark and Bergström, 2007, Tidermark, Bergström et al., 2003) and four of fair quality (Brazier, Walters et al., 1996, Coast, Peters et al., 1998, Holland, Smith et al., 2004, Lung, Howard et al., 2017). Three of these studies were carried out in fracture samples in Sweden (Parsons, Griffin et al., 2014, Tidermark and Bergström, 2007, Tidermark, Bergström et al., 2003), one in cognitively impaired older people in Germany (Kunz, 2010), two in recent hospital discharges (Coast, Peters et al., 1998, Holland, Smith et al., 2004) and one in a nursing home population (Lung, Howard et al., 2017) and one in a general group (Brazier, Walters et al., 1996). The studies measured responsiveness using either effect sizes, standardised effect sizes and standardised response means or criterion or construct based assessments with change scores either compared in known groups which would be expected to differ in change in score over time or change scores compared and correlated between measures of the same or similar

constructs. The period over which the change was allowed to occur ranged between 4 weeks (Parsons, Griffin et al., 2014) and 1 year (Kunz, 2010) and it was often not clear what was hypothesised to happen to EQ-5D scores during these periods. The studies in dementia (Kunz, 2010) and nursing home samples (Lung, Howard et al., 2017) reported small effect sizes, while the recent hospital discharge and fracture studies reported moderate-large effect sizes. In the three hip fracture sample external criterions of those expected to have good and less good early clinical outcomes found significant differences in change scores between these groups, with large corresponding SES and SRMs (Tidermark, Bergström et al., 2003, Tidermark, Zethraeus et al., 2002). However, results using correlations between change scores from comparator measures showed less clear results with weak and moderate correlations with change scores from the SF-12 PCS and the SF-36 (Tidermark and Bergström, 2007, Tidermark, Zethraeus et al., 2002).

Information about response distributions, floor/ceiling effects and missing data were reported to varying levels. One common issue in the studies found was a lack of detail about missing response rates, both at the overall measure level as well as for each item. Some studies made no mention of missing data and how this was dealt with at all (Davis, Bryan et al., 2012, Liang and Wu, 2014, Lutomski, Krabbe et al., 2017, Parsons, Griffin et al., 2014, Sanchez-Arenas, Vargas-Alarcon et al., 2014, Tidermark, Zethraeus et al., 2002), one stated that they only analysed fully completed EQ-5D responses (Lung, Howard et al., 2017) and some provided only a completion rate (Holland, Smith et al., 2004, Tidermark and Bergström, 2007, Tidermark, Bergström et al., 2003), while other studies provided data on the number of responses received at the item level. Where item level missing data rates were reported, they were less than 10% of responses missing for each item (Brazier, Walters et al., 1996) and for the majority of these studies the rate was below 5% (Coast, Peters et al., 1998, Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Kaambwa, Gill et al., 2015, Kunz, 2010, van Leeuwen, Bosmans et al., 2015b). Where completion rates were provided these were 81% (Holland, Smith et al., 2004) and approximately 98% (Tidermark and Bergström, 2007, Tidermark, Bergström et al., 2003).

Floor and ceiling effects could be reported for EQ-5D utility scores or individual items. Floors and ceilings in utility scores were less commonly seen with three studies reporting ceiling effects, with 15% of social care users (Kaambwa, Gill et al., 2015), 19% of respondents (mix of conditions and healthy group) (Lutomski, Krabbe et al., 2017) and 22% of fracture patients (Tidermark and Bergström, 2007) responding

11111 on the EQ-5D. No floors in utility scores were reported by any study. Ceiling effects for the EQ-5D items were commonly reported. While each item displayed a ceiling effect in at least one study, ceilings were most commonly seen in anxiety and depression (Coast, Peters et al., 1998, Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Kaambwa, Gill et al., 2015, Kunz, 2010, Lutomski, Krabbe et al., 2017, Sanchez-Arenas, Vargas-Alarcon et al., 2014) and self-care (Kaambwa, Gill et al., 2015, Kunz, 2010, Lutomski, Krabbe et al., 2017, Sanchez-Arenas, Vargas-Alarcon et al., 2014). Item floor effects were also seen in some studies, but less often than ceiling effects. The usual activities item suffered from floor effects most often (Coast, Peters et al., 1998, Diaz-Redondo, Rodriguez-Blazquez et al., 2014, Liang and Wu, 2014). There were also cases where categories were underused. Five different studies reported that less than 5% of respondents selected the worst category from anxiety/depression (Coast, Peters et al., 1998, Hofman, Lutomski et al., 2017, Holland, Smith et al., 2004, Kaambwa, Gill et al., 2015, Lutomski, Krabbe et al., 2017) and mobility (Holland, Smith et al., 2004, Kaambwa, Gill et al., 2015, Kunz, 2010, Lutomski, Krabbe et al., 2017, Sanchez-Arenas, Vargas-Alarcon et al., 2014).

Only one study, set in Australia and rated as good quality, investigated the measurement properties of the EQ-5D-5L (Ratcliffe, Flint et al., 2017). This study was based on a sample of 240 frail older people living in residential care who had recently experienced a hospital stay due to hip fracture, from an RCT of multidisciplinary rehabilitation services for hip fracture patients. This study examined the construct validity of the EQ-5D-5L in terms of convergent validity with clinical indicator measures of cognition (Mini Mental State Examination), depression (Cornell Scale for Depression in Dementia), pain (Pain Assessment in Advanced Dementia Scale) and functioning (Modified Barthel Index). Known group validity was also assessed in terms of the EQ-5D-5L's ability to statistically significantly distinguish between groups of participants either side of recommended threshold levels and severity categorisations on the clinical indicator measures of pain, depression and functioning. At the baseline measurement the study found significant but weak correlations between the EQ-5D-5L and the MMSE and the CSDD and a significant moderate correlation between the EQ-5D-5L and the PainAd, however by the 4-week measurement these correlations were all insignificant and very weak. At baseline, tests of known group validity showed that the EQ-5D-5L could significantly discriminate those individuals in different severity categories on the depression, pain and functioning scales in the expected directions. Effect sizes were significant but weak-moderate. However again by week 4 these relationships were no longer significant.

To summarise, there was strong evidence of the construct validity of the EQ-5D-3L in older people and moderate evidence of responsiveness. However, there was much more limited and conflicting evidence of the structural validity, internal consistency and reliability (test-retest and inter-rater) of the EQ-5D-3L in older adults. While evidence of the content validity of the EQ-5D-3L in older adults was judged to be strong, according to the synthesis criteria of one excellent quality study, this study was conducted in the Netherlands on the Dutch version of the EQ-5D-3L and therefore with translation changes and differences in cultural attitudes towards health and QoL, these results may not be generalizable to the UK setting (Fayers and Machin, 2016), leaving no evidence of the content validity of the English version of the EQ-5D-3L. Only one study was found to assess any element of psychometric performance of the EQ-5D-5L in older adults, which examined only construct validity. All other elements of its psychometric performance in older people remain in question.

SF-12

The measurement properties of the SF-12 in older adults were assessed in nine studies, reported in ten papers. All these studies considered the measurement performance of the SF-12 in general populations of older adults but two studies also considered specific groups alongside a general population group. One such study included a small stroke group (Jakobsson, Westergren et al., 2012) while the other included a group of older patients recently discharged from acute hospital inpatient stays (Resnick and Parker, 2001). Six studies included only community dwelling older adults, two included any living situation (Jakobsson, 2007, Jakobsson, Westergren et al., 2012) and two were unclear (Liang and Wu, 2014, Resnick and Parker, 2001). One study examining the measurement performance of the SF-12 was conducted in the UK (Pettit, Livingston et al., 2001), while four were undertaken in the USA (Cernin, Cresci et al., 2010, Fleishman and Lawrence, 2003, Resnick and Nahm, 2001, Resnick and Parker, 2001), two in China (Liang and Wu, 2014, Shou, Ren et al., 2016) and one each in Israel (Bentur and King, 2010) and Sweden (Jakobsson, 2007, Jakobsson, Westergren et al., 2012). The minimum age of participants in these studies ranged from 60 years old (Cernin, Cresci et al., 2010, Liang and Wu, 2014) to 75 (Jakobsson, 2007, Jakobsson, Westergren et al., 2012).

As outlined in the measure development section 3.4.5 of this thesis, changes have been made to the SF-12 at several points meaning there are several versions. SF-12 version 2 is the most up to date. The most significant changes are that, rather than dichotomous “yes/no” responses to the physical and mental role domains in version

1, these four questions now have 5-response Likert scales ranging from “all of the time” to “none of the time”. Also, the 6-response Likert scales for the vitality, mental health and social functioning items from version one were altered to 5-response Likert scales in version 2. It could be deduced from five of the eight studies investigating psychometric properties of the SF-12 in older adults that version one had been used (Bentur and King, 2010, Cernin, Cresci et al., 2010, Liang and Wu, 2014, Resnick and Nahm, 2001, Resnick and Parker, 2001). It was not clear from the remaining four which version had been used as a version number was not stated and no response distributions were provided from which this information could be deduced. Therefore, we cannot be sure that any of the psychometric analysis presented for the SF-12 reflects the changes made to this measure.

There is strong evidence of the construct validity of the SF-12, which was assessed in seven studies, discussed in eight papers, all of which used CTT methods. Four studies were judged to be of good quality (Cernin, Cresci et al., 2010, Pettit, Livingston et al., 2001, Resnick and Nahm, 2001, Resnick and Parker, 2001) and three fair quality (Bentur and King, 2010, Jakobsson, 2007, Jakobsson, Westergren et al., 2012, Shou, Ren et al., 2016). One study examining the construct validity of the SF-12 was conducted in the UK (Pettit, Livingston et al., 2001), while three were undertaken in the USA (Cernin, Cresci et al., 2010, Resnick and Nahm, 2001, Resnick and Parker, 2001) and one each in China (Shou, Ren et al., 2016), Israel (Bentur and King, 2010) and Sweden (Jakobsson, 2007, Jakobsson, Westergren et al., 2012). All seven studies focussed on the general older population while one study also included a condition specific stroke group (Jakobsson, Westergren et al., 2012) and one included a group recently discharged from hospital (Resnick and Parker, 2001). Two of the general population studies were ethnicity specific and focussed on an African American (Cernin, Cresci et al., 2010) and a Chinese (Shou, Ren et al., 2016) population.

Five studies investigated the known group validity of the SF-12 (Cernin, Cresci et al., 2010, Pettit, Livingston et al., 2001, Resnick and Nahm, 2001, Resnick and Parker, 2001, Shou, Ren et al., 2016). Characteristics on which tests of known group validity were based were: demographic characteristics such as age, gender and education level; the presence of self-reported long-term conditions and comorbidities; measures of recent service use such as GP visits, home care services and hospital inpatient/outpatient/A and E attendances; activity level; and whether the SF-12 could distinguish between those with and without various conditions such as depression

and dementia. Both subscales of the SF-12 were found to be able to significantly distinguish between respondents based on age, educational level and economic status (Shou, Ren et al., 2016) as well as those with and without LTCs (Cernin, Cresci et al., 2010, Resnick and Nahm, 2001, Resnick and Parker, 2001, Shou, Ren et al., 2016), self-reported health problems (Pettit, Livingston et al., 2001), ADL limitation (Pettit, Livingston et al., 2001) and recent service use (except A and E attendance and nursing home days) (Cernin, Cresci et al., 2010). Both subscales could also significantly distinguish between those taking more prescription medications (Cernin, Cresci et al., 2010) and those with depression and vision problems (Pettit, Livingston et al., 2001). Two studies also found that both subscales could significantly distinguish between those who did regular activity (Cernin, Cresci et al., 2010, Resnick and Nahm, 2001) while another found that only the physical subscale could significantly discriminate between these groups (Resnick and Parker, 2001). The physical subscale was the only one which could significantly discriminate between those with and without hearing problems and dementia, while the mental subscale was the only one which could significantly distinguish between those who did and did not report psychiatric problems (Pettit, Livingston et al., 2001). Neither subscale could significantly discriminate between gender and marital status (Shou, Ren et al., 2016).

Three studies measured the convergent validity of the SF-12 (Bentur and King, 2010, Jakobsson, 2007, Jakobsson, Westergren et al., 2012, Pettit, Livingston et al., 2001). All three included measures of activities of daily living as comparators as well as measures related to mental issues such as depression and anxiety. One also included a diagnostic scale for dementia (Pettit, Livingston et al., 2001) while another included questions about pain and mobility issues (Jakobsson, 2007, Jakobsson, Westergren et al., 2012). Two studies used correlations between the physical and mental subscales and scores on other measures to assess convergent validity (Bentur and King, 2010, Jakobsson, 2007, Jakobsson, Westergren et al., 2012) while the remaining study used forward linear regression methods to assess how much variation in the score of the other measure was accounted for by the physical and mental subscales of the SF-12 (Pettit, Livingston et al., 2001). Measures of ADL limitation, pain and mobility issues tended to be moderately – strongly correlated with the PCS while measures of depression and nervousness/worry were moderately – strongly correlated to the MCS. This was mirrored in the third study using regression methods.

There is conflicting evidence of the structural validity of the SF-12 in older adults, which was also assessed in seven studies (Bentur and King, 2010, Cernin, Cresci et al., 2010, Fleishman and Lawrence, 2003, Jakobsson, 2007, Jakobsson, Westergren et al., 2012, Resnick and Nahm, 2001, Resnick and Parker, 2001, Shou, Ren et al., 2016), using a variety of factor analytic methods such as EFA, CFA and structural equation modelling. Four of these studies were conducted in the USA (Cernin, Cresci et al., 2010, Fleishman and Lawrence, 2003, Resnick and Nahm, 2001, Resnick and Parker, 2001), of which two were rated good (Cernin, Cresci et al., 2010, Resnick and Nahm, 2001), one excellent (Fleishman and Lawrence, 2003) and one fair (Resnick and Parker, 2001), while the remaining three were conducted in non-English language translations of the SF-12 in Sweden (Jakobsson, 2007, Jakobsson, Westergren et al., 2012)(rated poor), Israel (Bentur and King, 2010)(rated poor) and China (Shou, Ren et al., 2016)(rated good).

The eight scales (made up of 12 items) of the SF-12v2 are hypothesised to form a two factor measure, with the physical functioning, physical role and pain scales hypothesised to have a strong association with the physical factor, the emotional role and mental health scales hypothesised to have a strong association with the mental factor and the general health, vitality and social functioning scales hypothesised to be moderately associated with both factors (with a stronger association between social functioning and mental health) (Maruish, 2012). This hypothesised structure has been supported by principal component analysis of the US 2009 and 1998 general population normative data (Maruish, 2012) .

However, factor structure results from the studies identified in this systematic review sometimes varied from the hypothesised structure above. The four USA based studies and the Chinese study found two factor structures (Cernin, Cresci et al., 2010, Fleishman and Lawrence, 2003, Resnick and Nahm, 2001, Resnick and Parker, 2001, Shou, Ren et al., 2016), while the remaining two studies of poor-quality reported extracting three factors (Jakobsson, 2007, Bentur and King, 2010). However, the make-up of these three factors varied greatly between the two studies and did not necessarily make additional theoretical sense and are not discussed further. While results varied between the US and Chinese studies, some clear patterns emerged. Physical functioning and physical role loaded onto the physical factor only, while the mental health scale loaded onto the mental factor only in all five studies. In the four US studies pain loaded on the physical factor and emotional role on the mental factor, as hypothesised. However, in the Chinese study emotional role was associated

strongly with the physical factor while pain loaded on both factors (Shou, Ren et al., 2016) which is inconsistent with the hypothesised structure of the measure. The Fleishman and Lawrence study (Fleishman and Lawrence, 2003) was the only one of the five studies to find the hypothesised split loading for general health, which was found to load solely on the physical factors in the remaining studies. All five studies found split loadings for social functioning, although this item tended to load stronger onto physical health rather than mental health as hypothesised. Two of the five also found split loadings for vitality (Cernin, Cresci et al., 2010, Fleishman and Lawrence, 2003), while in the remaining two US studies this item loaded onto the physical factor only (Resnick and Nahm, 2001, Resnick and Parker, 2001) and in the Chinese study vitality loaded only on the mental factor (Shou, Ren et al., 2016). Therefore, while the hypothesised factor structure was fairly consistent across the US studies (with a few minor variations for the scales which were hypothesised to associate moderately with both factors), bigger inconsistencies with the hypothesised structure were seen in the international studies.

One study assessed the presence of differential item functioning in relation to a series of demographic variables such as age, gender, education and ethnicity (Fleishman and Lawrence, 2003) using a multiple indicator multiple variance model (MIMIC). In relation to age, the presence of direct DiF was found on some items. Older adults tended to rate themselves more highly on the vitality, mental health and social functioning domains and lower on the physical functioning domain than would be expected from their underlying physical and mental health. It was also found that without adjusting for DiF, amongst the older age groups, mental health increased in the oldest age group, however once DIF was adjusted for the effect reversed showing lower scores in older groups as would be expected.

Strong evidence of the internal consistency reliability of the SF-12 was found. This property was assessed in seven studies, of which five were of good quality (Cernin, Cresci et al., 2010, Jakobsson, 2007, Jakobsson, Westergren et al., 2012, Liang and Wu, 2014, Resnick and Nahm, 2001, Resnick and Parker, 2001), one of fair quality (Bentur and King, 2010) and one of poor quality (Shou, Ren et al., 2016). One study only reported the Cronbach's alpha of the SF-12 as a whole despite having conducted factor analysis which revealed more than one factor (Shou, Ren et al., 2016). One study found 3 separate factors, with the physical role questions separating themselves from the physical factor and therefore reported the Cronbach's alpha of each of these domains separately (Bentur and King, 2010). The remaining five studies

found two factor models and reported the Cronbach's alpha for separately for the physical and mental subscales. These alphas ranged from 0.45-0.87 for the physical scale and 0.76 -0.80 for the mental scale. Only one study reported a Cronbach's alpha below 0.7 (0.45 PCS) (Cernin, Cresci et al., 2010) and none reported an alpha greater than 0.95 suggesting redundancy.

Limited evidence of test-retest reliability was found. This property was assessed in one US study which was judged to be of fair quality in relation to this measurement property (Resnick and Parker, 2001). Test-retest reliability was tested in older adults from the general population group through a repeat interview, 2-4 weeks after the initial interview. Correlations between the physical and mental subscale scores at the two time points were strong and significant at 0.86 for the PCS and 0.73 for the MCS.

The level of detail provided about missing data rates and response distributions was mixed. Two studies provided overall completion rates, both of which were over 94%, but no item level missing rates (Pettit, Livingston et al., 2001, Shou, Ren et al., 2016). Two further studies found 14% of respondents missed at least one item (Fleishman and Lawrence, 2003, Jakobsson, 2007, Jakobsson, Westergren et al., 2012), with one describing that the items with the highest rates of missing data were emotional role accomplish less at 5.9% missing and emotional role less carefully 7.9% (Jakobsson, 2007, Jakobsson, Westergren et al., 2012). The remaining five studies made no mention of missing data rates (Bentur and King, 2010, Cernin, Cresci et al., 2010, Liang and Wu, 2014, Resnick and Nahm, 2001, Resnick and Parker, 2001). Only two studies provided information about response distributions (Bentur and King, 2010, Liang and Wu, 2014). The Bentur et al study reported ceiling effects for the physical role questions, emotional role questions and social functioning. The Liang study found floor effects for all items (Bentur and King, 2010).

In summary there was strong evidence of the construct validity and internal consistency of the SF-12. However, there was limited evidence of its test-retest reliability and conflicting evidence of its structural validity. There was a pattern that US studies tended to obtain factor structures which mostly corresponded with the hypothesised structure of the SF-12v2, with only slight variation in the items which are hypothesised to load onto both factors. However, there were broader inconsistencies in the factor structures obtained from studies outside of the USA. No studies were found investigating the content validity, inter-rater reliability or responsiveness of the SF-12 in older adults. It is also unclear whether any validation of the newest version

of the SF-12, the SF-12v 2 has been conducted in older adults as all studies found either used version 1 or the version used is unclear.

ASCOT

Measurement properties of the ASCOT in older adults were tested in four studies, described in seven papers. Two of these studies were set in the UK (Hackert, Exel et al., 2017, Malley, Towers et al., 2012, Netten, Burge et al., 2012), one in the Netherlands (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b, van Leeuwen, Jansen et al., 2015) and one in Australia (Kaambwa, Gill et al., 2015). The study based in the Netherlands was undertaken in participants of a RCT evaluation of a geriatric care model for frail older adults living at home (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b, van Leeuwen, Jansen et al., 2015), while the UK studies were conducted in older social care service users (Hackert, Exel et al., 2017, Malley, Towers et al., 2012, Netten, Burge et al., 2012) and the Australian study included community dwelling older people receiving aged care services. The minimum age of participants in these studies ranged from 65 years old (Kaambwa, Gill et al., 2015, Malley, Towers et al., 2012, Netten, Burge et al., 2012, van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b, van Leeuwen, Jansen et al., 2015) to 75 (Hackert, Exel et al., 2017).

The focus of the evaluations tended to be construct validity, with all four studies examining this property in some way, resulting in strong evidence for this property. All four assessed convergent validity and two assessed known group validity (Kaambwa, Gill et al., 2015, van Leeuwen, Malley et al., 2014). Convergent validity was often examined between the ASCOT and other measures of similar constructs of health (EQ-5D-3L (Kaambwa, Gill et al., 2015, van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b), EQ-5D-5L, GDS-15 (Hackert, Exel et al., 2017), GHQ-12 (Malley, Towers et al., 2012, Netten, Burge et al., 2012), SF-12 PCS and MCS, Global Health rating scale (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b)); activities of daily living (ADLs (Malley, Towers et al., 2012, Netten, Burge et al., 2012, van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b), IADLs (Malley, Towers et al., 2012, Netten, Burge et al., 2012), Barthel Index (Kaambwa, Gill et al., 2015)); and broader concepts of QoL and wellbeing (OPQOL-13, SWLS, Cantrils Ladder (Hackert, Exel et al., 2017), OPQOL-Brief (Kaambwa, Gill et al., 2015), CASP autonomy subscale, UCLA Loneliness scale

(Malley, Towers et al., 2012, Netten, Burge et al., 2012), Global QoL rating scale (Malley, Towers et al., 2012, Netten, Burge et al., 2012, van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b), Pearlin Mastery Scale (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b) and measures of environment and service quality (Malley, Towers et al., 2012, Netten, Burge et al., 2012, van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b). Strong correlations tended to be found between ASCOT scores and scores of broader measures of QoL and wellbeing and moderate correlations between the ASCOT and health measures and measures of activities of daily living. These moderate correlations tended to be stronger in health measures with a mental focus and weaker in measures with more of a physical focus. The only items for which studies struggled to find evidence of construct validity were food/drink and dignity (Malley, Towers et al., 2012, Netten, Burge et al., 2012, van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b).

Known group validity was assessed in two studies; one of good quality (Kaambwa, Gill et al., 2015) and one of fair quality (Hackert, Exel et al., 2017). The ability of the ASCOT to discriminate between known groups of respondents who reported differing levels of health (Hackert, Exel et al., 2017, Kaambwa, Gill et al., 2015) and wellbeing (Hackert, Exel et al., 2017) as well as groups with different characteristics such as age, gender, education, informal care receipt and living arrangements (Kaambwa, Gill et al., 2015) was assessed. The Kaambwa study found that the ASCOT could significantly differentiate between those with differing levels of self-reported health in the expected direction but none of the other sociodemographic variables, living arrangements or informal care receipt resulting in significant differences (Kaambwa, Gill et al., 2015). The Hackert study reported that on average ASCOT scores were higher in those with above average health and wellbeing but it was not clear whether these relationships were significant (Hackert, Exel et al., 2017). Interestingly both studies found that ASCOT scores were higher in the higher age groups of older people, however neither study provided evidence of a significant relationship here.

One study from the UK also assessed the structural validity of the ASCOT in older adults using EFA but the model also included the ICECAP-O (Hackert, Exel et al., 2017). When taken together with the ICECAP-O items the ASCOT items split across three factors however, we cannot know how the addition of another measure, which may not measure the same underlying construct, impacted the reported dimensionality of the ASCOT. Therefore, this property was judged as still unknown in

the synthesis. More research is therefore needed to examine the dimensionality of the ASCOT itself in older adults.

Moderate evidence of the content validity of the ASCOT in older adults was found as it was examined in one Dutch study (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Jansen et al., 2015). Content validity was found to be adequate, with items largely understood as intended by the measure developers and found to be relevant to older respondent's QoL. There were some interpretation issues for the items; safety, control and dignity. These were thought to be due to language differences and their wording was subsequently altered. People found the dignity question confusing as they didn't understand how support and care would influence the way they thought about themselves. This was reflected in a higher rate of missing responses for this question than the others (10% vs 2%). There were also some issues noted with response options, with some respondents struggling to distinguish between the top two levels for occupation and misunderstanding some of the options for food/drink. It was also found that some respondents had difficulty in selecting a single response option for the social question as the options contain several elements, only some of which applied to the respondent's situation.

Limited evidence of the test-retest reliability and responsiveness of the ASCOT in older adults was also found. These properties were examined in one Dutch study (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b). Test retest reliability, assessed over a period of 1-2 weeks, was found to be good with an ICC (95% CI) = 0.71 (0.60, 0.78), although the confidence interval does cross the cut-off for good reliability. Responsiveness was measured using correlations between change scores from the same measures used to assess convergent validity. Correlations were found to be weak.

A consistent finding across studies was the presence of substantial ceiling effects for many ASCOT items. This was a particular issue for the items: personal cleanliness and comfort, accommodation cleanliness and comfort, food/drink and safety and often dignity and social participation. There was also often also a lack of respondents choosing the lowest response option for many items with the lowest category for each item tending to be chosen by less than 5% of respondents. Two of the studies reported missing data rates of 0% for the ASCOT (Hackert, Exel et al., 2017, Kaambwa, Gill et al., 2015). One of these studies was an online study in which all question had to be answered in order to move on (Hackert, Exel et al., 2017). The other UK validation of

the ASCOT reported item level rates of missingness between 10.3% for control and 9.3% for personal cleanliness/comfort, safety and dignity (Malley, Towers et al., 2012, Netten, Burge et al., 2012). The Dutch study reported a rate of missingness for the ASCOT index score of 14.7%, stating that this was mostly due to missing responses to the dignity item, which was missed by 12.6% of respondents (van Leeuwen, Bosmans et al., 2015a, van Leeuwen, Bosmans et al., 2015b).

There was strong evidence for the construct validity of the ASCOT in older adults and moderate evidence of its content validity. However, again this evidence of content validity was based on the Dutch version and therefore may not be fully generalisable to the UK version and older population. There was limited evidence of the responsiveness and test-retest reliability of the ASCOT. One study was found to examine the structural validity of the ASCOT however, this was in combination with the ICECAP-O and therefore the structural validity of the ASCOT itself remains unclear. No evidence was found for the internal consistency or inter-rater reliability of the ASCOT in older adults.

ONS-4 and WEMWBS

No studies were found investigating the measurement properties of the ONS-4 or WEMWBS in older adults. This is a large research gap which needs to be investigated to ensure that these wellbeing measures are valid, reliable and acceptable in older adults.

3.5.5 Discussion

3.5.5.1 Key findings

There was considerable variability in the amount and quality of psychometric evidence available for the different measures investigated. Unsurprisingly, most evidence was available for the EQ-5D-3L and SF-12 as these measures are much older and their use is widely established. Moreover, the EQ-5D-3L is the measure preferred by NICE in economic evaluation in England and therefore it is the most widely used measure of HRQoL in the UK and possibly in other countries.

While there was strong evidence of the construct validity of the EQ-5D-3L in older adults and moderate evidence of responsiveness, there was much more limited evidence of the structural validity and internal consistency of the EQ-5D-3L in older adults and limited and conflicting evidence of reliability (test-retest and inter-rater). While the evidence of the content validity of the EQ-5D-3L in older adults was judged to be moderate, this study was conducted on the Dutch version of the EQ-5D-3L in the Netherlands (van Leeuwen, Jansen et al., 2015) and therefore with translation changes and differences in cultural attitudes towards health and QoL, these results may not be generalizable to the UK setting (Fayers and Machin, 2016), leaving no evidence of the content validity of the English version of the EQ-5D. Only one study was found to assess any element of psychometric performance of the EQ-5D-5L in older adults, which examined only construct validity. All other elements of its psychometric performance in older adults remain in question.

For the SF-12 there was strong evidence of the construct validity and internal consistency of the measure. There was limited evidence of test-retest reliability and conflicting evidence on the structural validity of the SF-12, with US studies obtaining factor structures which mostly corresponded with the hypothesised structure of the SF-12v2, while studies outside of the US identified factor structures which were broadly inconsistent with the hypothesised structure.. No studies were found investigating the content validity, inter-rater reliability or responsiveness of the SF-12 in older people. It is also unclear whether any validation of the SF-12v2 has been conducted in older adults as all studies found either used version 1 or the version used is unclear.

There was strong evidence for the construct validity of the ASCOT in older adults and moderate evidence of its content validity. However, again this evidence of content validity was from the Dutch version (van Leeuwen, Jansen et al., 2015) and therefore may not be fully generalisable to the UK version and older population. There was limited evidence of the responsiveness and test-retest reliability of the ASCOT and, despite one study examining its structural validity, this was in combination with the ICECAP-O and therefore the structural validity of the ASCOT remains unknown. There was no evidence found for the internal consistency or inter-rater reliability of the ASCOT in older adults.

An important gap is evident from the findings of this review, as no studies were found to assess the psychometric properties of either the WEMWBS or the ONS-4 in older

adults. This is particularly important as interest in the use of wellbeing measures is increasing in economic evaluation. The NICE guidance on the economic evaluation of social care interventions state that wellbeing measures may be appropriate for assessing the benefits of such interventions on service users. A large proportion of social care users are over the age of 65. However, this review would suggest that there is a lack of evidence on the psychometric measurement performance of such measures in older adults. It is important that wellbeing measures such as the WEMWBS and ONS-4, which are currently being used in evaluations, are suitable in the population in which they are being used. More evidence is certainly needed on their psychometric performance in older adults.

Across the three instruments for which some evidence was found, there is substantial variation in the amount of evidence found for different psychometric properties. While we may be satisfied with the moderate to strong evidence found for construct validity, this evidence was all based on convergent and known-group validity. Only one study in the entire review was found which investigated DiF, for the SF-12 in the USA (Fleishman and Lawrence, 2003). Moderate evidence of content validity was also found for the EQ-5D-3L and ASCOT, however as already outlined, this was based on one study conducted in the Netherlands on the Dutch versions of these measures (van Leeuwen, Jansen et al., 2015). These findings may not be generalizable to the English versions or older population and therefore we cannot be sure about the construct validity in terms of measurement invariance or content validity of these measures in UK older adults. The evidence for other psychometric properties was mostly either limited, conflicting or none was identified suggesting we cannot be sure from the existing evidence about the performance of these measures in older adults in terms of structural validity, internal consistency, test-retest reliability, inter-rater reliability and responsiveness. More evidence on the performance of these measures in older adults in relation to any of the psychometric properties, except maybe convergent and known-group validity for the EQ-5D, SF-12 and ASCOT, would be of value.

One thing that was interesting to see from studies included in this review is the variety of age cut-offs used to define older adults. Minimum age cut-offs ranged from 60-80. This will create substantial difference in the resulting populations and potentially in the results of studies. This diversity in cut-offs used reflects a general lack of consensus about the age which it becomes appropriate to classify adults as “older adults”. Not only has there been great variation and debate about this, and little

agreement and consistency reached, but as time goes on and we see life expectancy and retirement ages rise it is likely that the appropriate cut-off, even if one was settled upon, will shift upwards over time. Careful consideration is needed as to the cut-offs chosen by studies and reasoning should be made clear as the cut-off chosen could have a large impact on the characteristics of the sample and the results.

It should be kept in mind that much of the evidence included in this review was conducted outside the UK, often on translated versions of the measures of interest. These results may not therefore be fully generalisable to the psychometric measurement performance of the measures in the UK. While for some measures there still may be sufficient UK based studies to remain confident that certain measurement properties are adequate in older adults, for some of the measurement properties there was no evidence at all for the English language version of the measure in a UK sample. For example, we may still be satisfied of the construct validity and responsiveness of the EQ-5D-3L, examined by four studies each. The remaining psychometric properties of the EQ-5D-3L are untested in a UK setting, except one assessment of test-retest reliability. For the SF-12 only one UK based study was found which investigated construct validity, leaving the rest of the psychometric properties unknown. Similarly, for the ASCOT two UK based studies were found. These both assessed construct validity, leaving a lack of evidence about the remaining properties. It is important to consider this, as the results of non-UK studies may not be fully generalisable to the UK older population, leaving a much more limited evidence base of the psychometric performance of these measures in older adults in the UK.

Studies largely focused on CTT methods for assessing psychometric properties, with the exception of structural validity which was assessed using factor analytic methods and one study which used structural equation modelling to assess DiF. IRT methods offer some important alternative insights and improvements over CTT methods when assessing psychometric performance. IRT methods enable researchers to have a closer look at how response options are being used which can indicate problems with these such as focussing effects, misunderstanding of level labels or levels which are indistinct from neighbouring categories. IRT also improves on the assessment of internal consistency. CTT methods assume that internal consistency and standard error of measurement around patients' scores are constant, regardless of the individuals' amount of the latent trait, but precision of measurement is known to vary by trait level (Hays, Staquet et al., 1998). IRT provides estimates of internal

consistency and standard error of measurement which vary by trait level, enabling the researcher to understand over what range of underlying health, QoL or wellbeing the measure provides a precise measurement. IRT methods also allow for the assessment of DiF, which is an important aspect of structural validity. The presence of DiF means that the property of measurement invariance does not hold and that people's scores are not solely determined by their level of the trait but are also dependent on their demographic characteristics. This may cause bias in scores and in any resulting decisions based on those scores. It is therefore important to measure and account for any DiF present.

3.5.5.2 Limitations

Older adults may have been included in general population or condition specific validation studies of the measures of interest. However, particularly in general population validation studies the number of older adults relative to the number of young adults tends to be small. These studies usually do not separate out results according to age, as this is not the focus of the validation study. Therefore, while it is a shame to exclude valuable data from older adults, we cannot be sure whether the overall results of such studies are truly representative of the older population. Studies which included some older adults but did not provide separate validation results for this group, therefore had to be excluded.

There are a huge range of PROMs which are potentially suitable for assessing the QoL and wellbeing of older adults. While it would have been preferable to systematically review the psychometric evidence of a broader range of instruments, this was not possible due to the time constraints of the PhD. Therefore, a selection of PROMs had to be chosen. The decision of which to include was based on which generic measures were found, in the literature and through discussion with local evaluators, to be used in the evaluation of integrated health and social care services, as the assumption of their use would be that they reflect the appropriate outcomes of these public sectors. However, there are many other PROMs, both generic and non-generic which could be appropriate in this area. This review certainly does not claim that the measures included are the only measures currently available which may be potentially suitable to evaluate the effectiveness of health and social care services in older adults.

3.5.6 Conclusion

This section systematically reviewed the evidence on the psychometric performance of the EQ-5D, SF-12, ASCOT, WEMWBS and ONS-4 in assessing the health, QoL and wellbeing of older adults. Most notably there was a lack of evidence on the psychometric performance of the WEMWBS and ONS-4 in measuring the wellbeing of older adults. This is an important gap in the literature as these measures are being used to evaluate services aimed at older adults in the UK. While there was at least moderate evidence of the construct validity of the remaining three measures in terms of convergent and known group validity there is less evidence on construct validity in terms of DiF, structural validity, internal consistency, responsiveness and test-retest and inter-rater reliability of these measures in older adults. There is also a lack of key evidence on the content validity of these measures in UK studies using the English language versions of the measures in older adults.

The important gaps identified in the existing literature on the psychometric performance of these measures clearly identified areas where this thesis could contribute to existing knowledge. These research gaps provided important guidance on the aims, objectives and design of this thesis study. In the following sections the rationale for this thesis study is outlined, followed by the aims and objectives that will be met in response to the identified research gap and the study design chosen to meet those aims and objectives.

3.6 Rationale and research gap

There is increasing evidence that older adults have different requirements of health and care services and have different priorities over what is important to their QoL than the working age population (Bulamu, Kaambwa et al., 2015). Older age is associated with frailty, characterised by a slow and steady decline in health and functioning (Milte, Walker et al., 2014). This means that older adults often have increasingly complex needs, requiring mixes of health and social care services. These services often have important aims and impacts on patients outside of health improvement (Bulamu, Kaambwa et al., 2015, Makai, Brouwer et al., 2014, Milte, Walker et al., 2014, van Leeuwen, Jansen et al., 2015). Examples of such broader aspects of QoL and

wellbeing, which have been found to be important to the QoL of older adults and are often goals of health and care services aimed at older adults, are independence, dignity and social contact (Makai, Brouwer et al., 2014, van Leeuwen, Jansen et al., 2015).

Older adults aged 65 and over make up a substantial proportion of social care users, with 51% of total social care expenditure spent on the over 65s in 2013/14 (Health and social care information centre, 2014). It has been recognised in the literature and by NICE themselves, that the current health focused measurement of QoL, using the EQ-5D, may not sufficiently capture the broader elements of QoL which are impacted by many health and care services required by older adults (Makai, Brouwer et al., 2014, van Leeuwen, Jansen et al., 2015). Broader measures of QoL such as the ASCOT and wellbeing measures have been suggested to be potentially appropriate to assess the impact of social care interventions on service users (National Institute for Health and Care Excellence, 2016).

PROMs are used to measure the impact of treatments and services on the QoL of patients and service users. For accurate assessments to be made, the assumptions that the measure provides a valid, comprehensive, reliable and responsive assessment of those aspects of QoL that are affected by services and important to patients must hold. Otherwise important impacts of services will be missed, leading to those services being undervalued, appearing less cost-effective than they are in reality and the most effective services for patients may not be funded (Makai, Brouwer et al., 2014, van Leeuwen, Jansen et al., 2015). Many PROMs are developed and psychometrically tested in patients or members of the general population. However, these assessments tend to focus on those aged 18-64, despite the fact that those same measures also go on to be used in older adults.

Content validity is of key importance to the psychometric performance of PROMs. If important aspects of QoL are missing from the measure, it may undervalue the impact of services that make a difference to these aspects of QoL. It is important to check that measures extensively cover those attributes of QoL that are important to its target population (Fayers and Machin, 2007). It also needs to be checked that questions are understood in the way measure developers intend and that questions are considered appropriate and relevant by the respondents to avoid invalid responses, increased distress and high missing response rates due to disengagement with the questionnaire. Measures with high levels of non-response or invalid answers are not

useful to evaluators as we cannot fully understand the impact of services on respondent's QoL. This is why thorough investigation of the content validity of the PROM, in all populations in which it will be used, is of such importance. Structural and construct validity are also crucial to ensure that statistical inferences drawn from responses to the measure are accurate and unbiased. It is important that patient characteristics, other than their underlying QoL, do not systematically impact their answers in different ways, resulting in DiF, as this can lead to bias in scoring and any resulting resource allocation decisions.

Potential issues surrounding the content validity of currently used wellbeing measures in terms of their acceptability and relevance were found in several local evaluations, as described in section 3.2.4. These issues led to consideration of whether these measures, and other commonly used measures of QoL, had been sufficiently validated in older populations.

Findings from the systematic review in Chapter 3 show that there is limited evidence of the psychometric performance of included measures in older adults. Most notably, no studies were found investigating the psychometric performance of the WEMWBS and ONS-4 wellbeing measures in older adults. This is a large research gap which is important to address. Particularly as wellbeing measures are mentioned as potentially appropriate in the evaluation of social care services in the NICE guidelines and half of the social care budget is spent on older adults. For the EQ-5D, SF-12 and ASCOT, there was at least moderate evidence of the construct validity of the measures in terms of convergent and known group validity in older adults, however there was a notable lack of evidence on the remaining elements of psychometric performance with evidence often limited or conflicting. While one study of excellent quality was found examining the content validity of the EQ-5D-3L and ASCOT in older adults, this study was conducted on the Dutch versions of these measures, and therefore may not be generalisable to the UK version or population. As seen in Chapter 2, content validity in all populations expected to receive the PROM is key for measures to provide accurate and comprehensive estimates of QoL. There was also a lack of evidence based on modern psychometric theories, such as IRT methods, which offer important advantages over popular CTT methods, including the examination of DiF, detailed information about how respondents use item response levels and estimates of internal consistency which vary across the underlying trait level. With these results in mind the aims and objectives of this thesis were formulated to address some of the important gaps in the literature.

The preferred QoL measure in social care guidelines is not currently settled (National Institute for Health and Care Excellence, 2016). This research will help inform the selection of a best performing QoL measures for both integrated health and social care services as well as social care alone, out of currently popular measures, as the elderly represent a substantial proportion of those currently requiring, and at high risk of soon requiring, social care as well as integrated services. While identifying measures which do and do not perform well in this group is important in itself to effectively evaluate interventions aiming to improve the QoL of frail elderly, moving towards the selection of a preferred measure would also enable more comparability between evaluations. Both of these aspects are important for unbiased resource allocation.

The important gaps identified in the existing literature on the psychometric performance of these measures provided important guidance on the aims, objectives and design of this thesis study. By providing evidence on such gaps this thesis seeks to contribute to existing knowledge as described in this chapter.

3.7 Study aims and objectives

The main aim of this thesis is to assess the psychometric performance of a selection of existing health, QoL and wellbeing measures in older adults, in order to explore whether they are suitable for use in the economic evaluation of health and social care services aimed at older adults. To meet this aim, the following objectives will be addressed:

- To examine the structural and construct validity, internal reliability and acceptability of the EQ-5D-5L, SF-12v2, ONS-4, ASCOT and WEMWBS in assessing the health, QoL and wellbeing of older people using item response theory and differential item functioning methods
- To examine the content validity of the EQ-5D-5L, SF-12v2, ONS-4 and WEMWBS in assessing the health, QoL and wellbeing of older people using cognitive interviews
- To provide information on which of the measures tested are appropriate to use in the evaluation of aged health and social care services.

3.8 Study design

As seen in previous chapters, different aspects of psychometric performance are explored using different methods. While most measurement properties can be examined using quantitative methods; such as structural and construct validity, reliability and responsiveness; the investigation of content validity requires the use of qualitative methods. Qualitative methods can also be useful to further examine issues identified using statistical methods by delving deeper into respondent's perceptions to find out why issues are occurring and suggest suitable solutions. Therefore, this PhD is made up of two studies, each aiming to examine different elements of psychometric performance of the selected measures in older adults. An overview of their design and how they are each expected to contribute towards the aims and objectives of the thesis is outlined in the following section. More detailed accounts of the methods used in each study are reported in the specific chapters relating to each study.

Study 1: An investigation into the psychometric performance of the EQ-5D-5L, SF-12v2, ASCOT, ONS-4 and WEMWBS in older adults using item response theory methods

The objective of Study 1 is to explore the structural and construct validity, internal reliability and acceptability of the EQ-5D-5L, SF-12v2, ONS-4, ASCOT and WEMWBS in assessing the health, QoL and wellbeing of older adults. Response data for adults of all ages were taken from various large available UK datasets in different populations for each measure. Each dataset was then split into adults aged under and over 65. Structural validity in terms of the dimensionality of each measure was assessed using factor analysis. Then multiple group GRM IRT models of the relevant structure were run and structural and construct validity, performance of item response options and internal reliability were compared between age groups and DiF was assessed.

IRT methods were chosen for this study as the literature review found very few studies had adopted these methods despite them having important advantages over CTT methods including: providing greater detail about the performance of items and response levels within them; allowing the SEM and therefore the estimate of internal

consistency reliability to vary between individuals with different levels of the underlying trait and allowing the examination of age related DiF, which can lead to bias in scores due to participant characteristics other than their level of underlying trait. Further details about the methods used in this study and the results obtained can be found in Chapter 4.

Study 2: An investigation into the content validity of the EQ-5D-5L, SF-12v2, ONS-4 and WEMWBS in older adults using cognitive interviewing techniques

The objective of Study 2 was to explore the content validity of the EQ-5D-5L, SF-12v2, WEMWBS and ONS-4 in assessing the QoL and wellbeing of older adults. Semi-structured cognitive interviews were used in this study to examine peoples' understanding and interpretation of the questions within each measure, whether they thought that these questions were appropriate to ask and relevant to older adults and whether the questions comprehensively covered what was important to the QoL and wellbeing of older adults. Interviews covered two of the four measures of interest in varying combinations. Interviews began with questions about the definitions of QoL and wellbeing and what was needed in life for the participants to feel they had a good level of QoL and wellbeing. Next the participants were asked to complete the first measure, using think aloud techniques. Semi-structured interview questions and verbal probing techniques were then used to further explore participants understanding and interpretation of questions and their opinions on the relevance and appropriateness of questions and the comprehensiveness of the measure. This process was then repeated for the second measure. The interview closed with a discussion of which of the two measures the participant preferred as a measure of their QoL or wellbeing.

The use of semi-structured interviews enabled an in-depth exploration into people's opinions about the selected measures. The use of cognitive techniques of think aloud and verbal probing provided the opportunity to examine people's thought processes when responding to the measures allowing the researcher to identify response issues and explore in greater depth peoples' interpretation and opinions on the measures and questions within them. Further details on the methods used and the results of this study can be found in Chapter 5.

3.9 Conclusion

In this chapter, the methods by which the measures which were chosen for examination in this thesis were identified and selected were outlined. Then, the details of each measure and the process by which they were developed was described. Next, the existing literature on the evidence of their psychometric performance in older adults was systematically reviewed. Following this the research gap and rationale for the thesis was outlined, before the aims and objectives were presented. Lastly, the study design chosen to meet the aims and objectives were introduced. The next chapter presents the first of two studies investigating the psychometric performance of the chosen measures in older adults.

Chapter 4

An investigation into the psychometric performance of the EQ-5D-5L, SF-12v2, ASCOT, ONS-4 and WEMWBS in older adults using item response theory methods

4.1 Introduction

The previous chapter explained the methods used and the justification for the choice of measures included within this thesis. The aims and objectives of the thesis, the rationale for the study and the study design were also outlined. This chapter presents the detailed methods, and results of the first of two studies outlined in Chapter 3. This study aims to investigate the construct validity, structural validity and internal consistency reliability of existing commonly used measures of health, QoL and wellbeing in older adults. This chapter will establish whether there is age related DiF in the data that respondents provide. This is important as, if bias exists it could bias the results of any economic evaluations of health and social care services aimed at older adults in which the relevant measure is used. This could affect resource allocation decisions, meaning that the most cost-effective services may be undervalued and not funded, which lowers the total health gain which can be obtained by the health budget. If interventions more suited to different age groups are competing for funding resources, bias in estimates of QoL and the impact services have on the QoL of different age groups may bias funding away/towards certain age groups. A glossary of terms related to this chapter can be found at the end of the chapter in section 4.7.

4.2 Aim

To examine the structural and construct validity, internal reliability and acceptability of the EQ-5D-5L, SF-12v2, ONS-4, ASCOT and WEMWBS in assessing the health, QoL and wellbeing of older adults using IRT methods.

4.3 Methods

As described in section 2.5, there are several schools of psychometric methods commonly used to assess the performance of PROMs and the items within them. Broadly these are split between the more traditional CTT methods and the modern latent trait methods such as IRT and Rasch Modelling. This section begins with a discussion of these methods and justification for the choice of IRT. Then the methods adopted in this IRT analysis are explained in detail before the results are presented in section 4.4 and the methods and results are discussed in section 4.5.

4.3.1 Choice of psychometric theory

Classical Test Theory versus Item Response Theory

CTT mostly involves the analysis correlations and descriptive statistics, while IRT and Rasch are based on more sophisticated statistical methods. Because of this IRT and Rasch are able to go beyond simply providing descriptive information about items and measures by highlighting potential causes and areas for improvement (Petrillo, Cano et al., 2015). IRT is argued to provide a much more detailed description of the performance of each item on a PROM than CTT, as it models the relationship between the trait being measured and each individual item (Bjorner, Kosinski et al., 2003, Nguyen, Han et al., 2014). Second, the standard error of measurement around patients' scores and internal consistency of the scale are assumed to be constant in CTT, regardless of the individuals amount of the latent trait, but precision of measurement is known to vary by trait level (Hays, Staquet et al., 1998). In IRT precision of measurement and internal consistency reliability are related to the information level, which varies by trait level, meaning IRT can provide estimates of measurement precision specific to the score level (Bjorner, Kosinski et al., 2003). Moreover, IRT can assess each items' contribution to the total precision of measurement of the measure for the specific score range (Bjorner, Kosinski et al., 2003). Therefore, IRT provides important information about the ranges of trait over which the measure has good or inadequate levels of internal consistency reliability and precision of measurement. Lastly the IRT framework provides a sensitive framework for the investigation of DiF (Fayers and Machin, 2016). This provides a test of measurement invariance, an important element of construct validity. Due to the above advantages of IRT methods over CTT methods, IRT methods were adopted in this study.

4.3.2 Choice of age cut-off

The age at which adults begin to be described as “older adults” was found to vary substantially between the studies found in the systematic review presented in Chapter 3. As acknowledged in section 3.5.5 this diversity in cut-offs used reflects a general lack of consensus about the age which it becomes appropriate to classify adults as “older adults”. Not only has there been great variation and debate about this, and little agreement and consistency reached, but as time goes on and we see life expectancy and retirement ages rise it is likely that the appropriate cut-off, even if one was settled upon, will shift upwards over time. Choice of cut-off will likely have a substantial impact on the characteristics of the resulting sample as the prevalence of negative outcomes such as frailty and functional decline and resulting increased risk of outcomes such as disability, care home and hospital admission and death, which are all known to increase with age (Fried, Tangen et al., 2001, Rockwood, Mitnitski et al., 2006). Therefore, choice of age cut-off is important. This study chose a cut-off of older adults being those aged 65+. This is the cut-off which seems most broadly accepted in the literature on ageing (Age UK, 2018b, World Health Organization, 2002) and reflects current state pension age in the UK (Age UK, 2018a). However, this choice is made with the understanding that in the future, as life expectancy and state pension ages rise, higher cut-offs may be more appropriate. An alternative cut-off of 75+ is tested in sensitivity analysis towards the end of this chapter.

4.3.3 Data sources

IRT and DiF analyses require large numbers of respondents. Exact sample size requirements are debated but it is usually argued that upwards of 500 respondents are required for robust two-parameter IRT models with at least 200 per group (under 65s and over 65s) for DiF analysis (Edelen and Reeve, 2007). No large UK general population datasets, which included all measures of interest, were found so analysis was conducted separately for each measure, with measures being drawn from different datasets. The datasets used are outlined in Table 4 and further details are discussed below. Further details on the characteristics of respondents from each dataset can be found in section 4.4.1.

Table 4 - Summary of Data Sources

Dataset	Questionnaire	Measures	Population	Sample Size
Health improvement and patient outcomes survey (HIPO) 2013-14	EQ-5D-5L SF-12v2 ONS-4	HRQoL Health status	Recently discharged hospital inpatients	Aged 18-64 = 3,632 Aged 65+ = 2,719
Adult social care survey 2014-15	ASCOT	SCRQoL	Social care service users	Aged 18-64 = 27,256 Aged 65+ = 41,755
Health survey for England 2014	WEMWBS	Mental wellbeing	General population	Aged 18-64 = 5,801 Aged 65+ = 2,069

Health Improvement and Patient Outcomes Dataset (HIPO) (EQ-5D-5L, SF-12v2, ONS-4)

The HIPO dataset is a large patient dataset which collected health and wellbeing data, including the EQ-5D-5L, SF-12v2, and ONS-4, from inpatients recently discharged from Cardiff and Vale NHS Hospital Trust in 2013-14 via postal survey (Mukuria, Rowen et al., 2016). The survey was sent to patients aged 18 and over 6 weeks after discharge. Patients from most specialties were included however those with a primary mental health diagnosis were not included. The survey was linked with routine hospital data, providing a dataset with sociodemographic (age, gender), health (EQ-5D-5L, SF-12v2), wellbeing (ONS-4, subjective wellbeing VAS and single positive and negative wellbeing questions) and diagnosis data. 25,919 surveys were sent between September 2013 and January 2014 and 6,351 completed questionnaires were returned (25% response rate). 1,007 (16%) returned questionnaires with at least one response missing.

Health Survey for England (HSE) (WEMWBS)

The WEMWBS is included in 2014 wave of the HSE; an annual general population survey conducted by computer assisted home interview in England (Health and Social Care Information Centre and Department of Health, 2014). The survey asks a large number of questions about participant's health and wellbeing and information about long-term conditions. While most of the interview is computer assisted, with the interviewer asking questions and filling in responses, there is also a participant self-completion booklet on paper which includes the EQ-5D-3L, GHQ-12 and WEMWBS.

The HSE survey is designed to cover a representative sample of the general population living in private households in England (Health and Social Care Information Centre, 2014). Those in nursing or residential homes are not covered by this survey. HSE is not a panel survey (it does not survey the same people and link their responses each year) meaning it cannot be used to follow individual's health and wellbeing over time. The survey covers both adults (aged 16 and over) and children (aged 0-15). However, the health and wellbeing measures are asked only to adults and therefore only adult data is used in this study. The adult sample size for the 2014 wave of the HSE was 8,077 adults from 5,051 households (a household response rate of 62%).

Adult Social Care Survey (ASCS) (ASCOT)

ASCOT was found in the ASCS, an annual postal survey of all adult users of long-term support services funded or managed by the social services (Health and Social Care Information Centre, 2015). The ASCS is conducted by councils with social services responsibility, to get feedback from service users about their experience of the services they receive and how they are helping them to live safely and independently in their own homes (NHS Digital, 2015). The survey includes those still living in the community, in receipt of long-term social care services, and those living in nursing and residential homes. The data used in this study come from the 2014-2015 round of the ASCS. Questionnaires were sent in March 2015. Of a sample of 192,995 service users, 69,510 responded (response rate 36%).

4.3.4 IRT analyses

IRT analyses were run in MPlus version 7.4 (Muthén and Muthén, 1998-2015). Data was prepared in STATA version 14 (StataCorp., 2015) and then converted to MPlus where all IRT analysis was undertaken using the stata2mplus conversion command (Statistical Consulting Group, Institute for Digital Research and Education et al.). Prior to conducting any analysis, datasets were randomly split into a model development and a model validation dataset. This was carried out in STATA by generating a variable which assigned a random number between 0 and 1 to each respondent. Those who received values above 0.5 were assigned to the model development sample, and those who received 0.5 or less were assigned to the model validation sample. Differences between samples were tested for using the Two-sample

Wilcoxon rank-sum (Mann-Whitney) test for continuous variables and Pearson's Chi-square test for categorical variables. IRT analysis was carried out on the model development sample and then the stability of results was checked by rerunning the final DiF model in the validation sample.

4.3.3.1 Assumption checks

IRT models can be unidimensional, assuming all items relate to a single latent concept, or multidimensional with items representing one of several constructs. It is important that dimensionality is assessed and the correct type of IRT model is run to obtain valid results. The dimensionality and structural validity of the included measures was examined using categorical EFA and CFA to ensure that the correct type of IRT model was chosen. Factor analytic methods examine the correlations between a set of observed variables, in this case items, and attempts to describe the variability between these items in terms of a lower number of unobserved constructs called factors (Field, 2009). EFA is an exploratory approach within this technique, which seeks to explain the correlation between the items using a number of factors based solely on the observed correlation matrix. CFA is a confirmatory approach, which seeks to verify the appropriateness of a given theoretical measurement model, which outlines the hypothetical relationships between items and a set of factors based on the theory behind the measure or previous empirical research (Field, 2009).

First EFA was run, to explore the underlying dimensional structure of each instrument, suggested by the relationships between items in the data. Eigenvalues and scree plots were examined to establish the appropriate number of factors underlying each measure, according to commonly used predefined decision rules. Eigenvalues represent the relative share of total variance accounted for by each factor. The Kaiser rule suggests keeping all factors with eigenvalues ≥ 1 (Kaiser, 1960). MPlus also provides a visual representation of the eigenvalues on what are called scree plots, which plot the proportion of total variance (y-axis) explained by each factor (x-axis) with factors ordered according to the proportion of total variance they explain. The scree-test (Cattell, 1966) involves identifying on a scree-plot where there ceases to be a sharp decrease in the eigenvalues of subsequent factors. The number of true factors is said to end where a line drawn through all the points becomes linear and flat.

Then CFA was run, to confirm the structure of potentially appropriate models based on EFA results and compare the fit of these models using measures of absolute and relative model fit. Absolute model fit is examined in terms of the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). Cut-offs for good model fit are $CFI \geq 0.95$ and $RMSEA \leq 0.05$, with $RMSEA \leq 0.08$ considered fair (Yu, 2002). Relative model fit was examined by comparing AIC and BIC values between models, with lower values signify better fit. Where unidimensionality was not clear, sufficient unidimensionality was tested using a bifactor model (McDonald, 1999). Bifactor models allow each item to load onto the relevant factor suggested by the multi-factor CFA and onto a single global factor. If the items load substantially higher onto the global factor than the other relevant factor and the global factor explains substantially more of the common variance than the other factors, the measure can be claimed sufficiently unidimensional and a unidimensional IRT can be run. If the global factor does not dominate, a multidimensional IRT is run (Reise, Bonifay et al., 2013). There are no set cut-offs for the amount of common variance which should be explained by the global factor for it to be considered dominant (Reise, Bonifay et al., 2013). However, it has been suggested in the literature that 50% of common variance explained by the global factor should be considered the minimum for sufficient unidimensionality, while values closer to 75% would be much preferred (Reise, Bonifay et al., 2013).

IRT models assume local independence, meaning there is no additional systematic covariance between items beyond their given relationship to the underlying trait (Edelen and Reeve, 2007). Local dependence may arise in groups of items with similar content or which are physically grouped together on a measure. Local dependence may therefore signal item redundancy. Large modification indices of error covariances between items may suggest local dependence. Local dependence and item redundancy may also be suggested by groups of items of similar content or grouping having substantially higher discrimination parameters than the other items within a measure. Local dependence may be investigated by removing one of the items within a group suspected to exhibit local dependence and watching for a substantial change in the discrimination parameters of the other items within that group. A substantial change in the other discriminations suggest local dependence and item redundancy. One option is to remove one of the offending items at this stage however this takes away the opportunity to gather additional information about the performance of this item. Longer scales, of approximately 10 or more items are also

argued to be robust to local dependence and therefore leaving these items in should not substantially impact models with more items (Edelen and Reeve, 2007).

4.3.3.2 Model comparisons

This study uses polytomous ordinal IRT models as the items in all PROMs considered have more than two ordered response options. Different types of polytomous ordinal IRT models were fitted to the data and compared in terms of relative and absolute fit statistics. Types of model tested were Graded Response Models (GRMs), constrained GRMs, Partial Credit Models (PCM) and Generalised Partial Credit Models (GPCMs). These models and the assumptions they make have been previously summarised in Section 2.5.2.2, Table 1. The RSM, with its additional assumption of equal spacing between difficulty parameters as well as equal discriminations, was considered too restrictive for PROMs (Nguyen, Han et al., 2014) as there is no evidence of equal spacing between response options in terms of the amount of QoL required to be most likely respond in different categories nor any evidence that different items on a PROM are equally important to the QoL of participants. Therefore, this model was not tested.

Both one-parameter and two-parameter versions of GRM and PCM were fitted to test the assumption of equal discrimination across items. By assuming equal discrimination in one-parameter models (constrained GRM and PCM), we assume all items are equally well related to the underlying trait. Two-parameter models (GRM and GPCM) allow items to have different discrimination parameters to allow for the fact that items may perform differently. First a two-parameter GRM was run for each measure. This was then compared in terms of nested model fit to a constrained one-parameter GRM to test the assumption of equal discrimination across items. By assuming equal discrimination, we assume all items are equally well related to the underlying trait, wellbeing. One-parameter models constrain discrimination parameters to be equal for all items, indicating that they are all equally well related to the underlying trait, while two-parameter models allow items to have different discrimination parameters to allow for the fact that items may perform differently. Whichever version of GRM was shown to have superior fit was compared in terms of absolute fit to the version of PCM with the same number of parameters. A nominal categories model with the same number of parameters was also run and compared in terms of fit to the ordinal models. This tested whether there were any problems with

the ordering of the levels as nominal categories models do not force the intended ordering of response categories.

GRMs and constrained GRMs were estimated using full information maximum likelihood (ML) estimation with the logit link. Identical versions of each GRM model were also run using WLSMV estimation with theta parameterization as this provides absolute fit statistics with which to compare models, which are not provided under ML estimation which only provides tests of nested relative fit. The GPCM and nominal models were estimated using robust maximum likelihood (RML) estimation.

Absolute model fit is examined in terms of the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). Cut-offs for good model fit are as follows; CFI >0.95 and RMSEA<0.05 with RMSEA<0.08 considered fair (Yu, 2002). Relative nested model fit is tested by a rescaled likelihood ratio test which examines the difference in the log likelihoods multiplied by minus two (-2*LL) of the two model calibrations, distributed as a Chi-square with degrees of freedom equal to the difference in the number of estimated parameters in the two models (Nguyen, Han et al., 2014). The null hypothesis is that the more parsimonious model, e.g.one-parameter rather than two-parameter, fits best. Significant differences in the test suggest that alternative to the null hypothesis the model with more parameters provides better fit. The AIC and BIC criteria also provide evidence on relative nested model fit with lower values indicating better relative fit. Whichever model showed superior fit was taken forward for DiF analyses.

4.3.5 DiF analyses

In order to investigate DiF multiple group models are run to test whether items behave differently in younger (aged 18-64) and older adults (aged 65+), and therefore exhibit DiF. The best fitting model from the IRT phase was carried forward and converted into a multiple-group model to test whether the items exhibited DIF, and therefore behaved differently, between under and over 65s. This multiple group model separates the data into under and over 65s and estimates parameters separately in the two groups. A step-by-step process has been developed and widely used to analyse DiF (Milfont and Fischer, 2010, Putnick and Bornstein, 2016, Vandenberg and Lance, 2000), in which the model needs to be rerun under a series of sets of conditions and restrictions, simultaneously in the two groups of interest. The stages and related

conditions and restrictions involved in their model set up are outlined below in Table 5 and described in detail in the following paragraphs. The impacts of these restrictions/conditions will be tested by comparing nested model fit. WLSMV estimation with the probit link and theta parameterization was used for all DiF models. WLSMV was chosen as it does not assume the underlying data follows a normal distribution, it provides absolute fit statistics and easy assessment of the impact of the restrictions of the various stages of DiF analysis by using the DIFFTEST function comparison of nested model fit and it runs multiple-group multiple-factor models much quicker than some other estimators.

Table 5 – Stages of DiF Analysis

Stage	Testing DiF in	Model Identification Set up	Parameter Constraints Procedure
Factor structure (baseline two-group model)	Dimensionality/ model structure	Factor variance fixed at 1, factor mean fixed at 0 and residual variances fixed at 1 in both groups	Discrimination and difficulty parameters free to vary across groups.
Non-uniform DIF	Discrimination parameters	Factor variance fixed at 1 in under 65 group and freely estimated in over 65 group. Factor mean fixed at 0 and residual variances fixed at 1 in both groups	Constrain discrimination parameters to be equal across groups. Compare fit to baseline two-group factor structure model. If fit significantly worse examine MIs to see which discrimination is causing most local misfit and free that discrimination. Continue until non-uniform DiF model fit is insignificantly different to baseline two-group factor structure model fit.
Uniform DiF	Difficulty parameters	Factor mean fixed at 0 and factor variance fixed at 1 in under 65 group, but both freely estimated in over 65 group. Residual variances fixed at 1 in both groups	Start with final non-uniform DiF model. Constrain difficulty parameters to be equal across groups. Compare fit to final non-uniform DiF model. If fit significantly worse examine MIs to see which difficulty parameter is causing most local misfit and free that difficulty. Continue until Uniform DiF model fit is insignificantly different to final non-uniform DiF model fit.

Stage	Testing DiF in	Model Identification Set up	Parameter Constraints Procedure
Residual variance	Unstandardised residual variances	First stage run Residual Free A model (Factor mean fixed at 0 and factor variance fixed at 1 in under 65 group, but both freely estimated in over 65 group). Residual variances fixed to 1 in under 65 group. Residual variances for those items which did not exhibit non-uniform DiF free in over 65s). Second stage rerun final Uniform DiF model.	Start with final Uniform DiF model. Free residual variances for those items which did not exhibit non-uniform DiF in over 65s (Residual Free A model). Compare fit to final Uniform DiF model. If fit insignificantly different between models, then residual variances can be fixed to 1. If significant difference in fit between models examine MIs to free worst performing residual variance until models insignificantly different.
Factor variance and factor mean	Factor variance and factor mean	Factor mean fixed at 0 and factor variance fixed at 1 in under 65 group. Factor mean free and factor variance fixed to 1 in over 65 group. Residual variances fixed at 1 in both groups	Start with final residual variance model. Constrain factor variance in both groups. Compare fit to final residual model. If difference in model fit insignificant then factor variance can be fixed to be equal across groups. Factor means are free to differ between groups in this model. Check if the factor mean estimate in over 65s indicates a significant difference.

The investigation of DiF has multiple stages. First a baseline two-group model is run, and the factor structure is examined to see if there are structural differences in the model between under and over 65s. In this stage discrimination parameters and difficulty parameters are free to vary between groups. If this model results in acceptable levels of fit, it can be concluded that the structural model, in terms of factor structure, is equal across groups. The factor variance was fixed to one and the factor mean was fixed to zero for identification. Residual variances are not uniquely identified in the factor structure invariance model and were therefore constrained to equal one in both groups.

Then the measure is tested for non-uniform DiF by investigating whether the unstandardised discrimination parameters are the same in the different groups, signalling that items are equally well related to the underlying trait for under and over 65s (Putnick and Bornstein, 2016). For identification, in this model the factor mean was fixed at zero for both groups while the factor variance was fixed to one in the under 65s and freely estimated in the over 65s. Residual variances were constrained to equal one in both groups. Discrimination parameters were free to vary across items but were constrained to be the same in both groups. Difficulty parameters were also freely estimated and allowed to vary between groups. If constraining the discrimination parameters to be equal across groups results in significantly worse nested model fit as compared to the baseline two-group factor structure model, signified by a DIFFTEST p-value <0.05 , then DiF in terms of discrimination parameters can be concluded. Modification indices can be used to free the worst performing item discrimination parameters one by one until the DIFFTEST, compared to the baseline two-group factor structure model, results in an insignificant p-value.

Thirdly, uniform DiF is tested for, to see if items exhibit DIF in terms of their difficulty parameters. Factor mean and variance were fixed at zero and one respectively in under 65s for identification but were freely estimated in over 65s. A model with all thresholds for all items constrained to be equal across groups is compared to the final non-uniform DiF model, in which all or some discrimination parameters were left constrained to be equal between groups, in terms of nested fit. If this uniform DiF model is found to fit significantly worse than the final non-uniform DiF model, thresholds are released one by one, based on those judged to have the most impact on misfit by the modification indices, until an insignificant DIFFTEST p-value results. Those thresholds which need to be freed are argued to exhibit DIF in terms of their difficulty parameters.

Next, the equality of unstandardized residual variances across age groups is tested. A model with residual variances freely estimated, except in those items which were found to exhibit DIF in their discrimination parameters in the non-uniform DiF model stage, in over 65s and residual variances are fixed to one in under 65s is compared to a model in which the residual variances are all fixed to one using the DIFFTEST function. The rest of the parameters remain the same as in the last uniform DiF model. If constraining the residual variances of over 65s results in a significantly worse model fit, then the modification indices were used to signal the worst performing item and the residual variance on that item is freed. This process continues until an

insignificantly different model to the first residual invariance model is found. Items which remain constrained to be equal in terms of residual variances at the end of the process signal that the amount of item variance not accounted for by the factor was the same across groups.

Finally, the equality of factor variances and factor means across groups is tested by constraining the previously free factor variance in over 65s to equal one and be equal to the factor variance of under 65s. If, when compared using DIFFTEST to the final residual invariance model, this results in significantly worse fit it signals a difference in the variability of QoL or wellbeing between the age groups. In this model the factor mean is set to 0 in under 65s and is free in over 65s. An insignificant p-value on the estimate of the factor mean in over 65s indicates that the factor mean is insignificantly different between the two groups.

While this step by step process systematically indicates where significant DiF lies, it does not provide information about the magnitude and impact of DiF (Teresi, Ocepek-Welikson et al., 2007). Just because DiF in item parameters is significant does not mean that this difference is large, or of practical importance in terms of its impact on scores. The magnitude of DiF can be examined using expected item scores. Expected item scores are calculated by summing the weighted probability of each response (weighted by the response category value) at each level of underlying wellbeing. These expected scores are calculated separately for each group based on probabilities of responses from the final residual invariance model in which some discrimination and difficulty parameters may have been freed. Differences in expected scores can then provide an estimate of the impact of DiF at the item level. These expected scores can then be summed across items to provide a scale level estimate of the impact of DIF at each level of wellbeing. The EQ-5D and ASCOT are preference-based measures, with weightings attached to each response option. Published preference weighting tariffs include the utility decrements associated with each response option and these were used as the weightings in the estimation of expected scores for these two preference weighted measures.

It was also felt beneficial to estimate effect sizes to aid interpretation of whether differences in expected scores between groups were of practical importance. Expected score standardised differences (ESSD) for each item were estimated following the procedure outlined by Meade (Meade, 2010). This measure is described as an expected score version of Cohen's d (Cohen, 1988), an effect size used to

indicate standardised difference between two means. To estimate the ESSD for an item, expected scores are estimated for the alternative group (the smaller in size of the two groups), using both sets of item parameters (those identified for under and over 65s). Then the mean expected score for this group under each set of parameters is estimated along with the standard deviations. Mean expected scores were calculated by exporting factor scores for each member of the alternative group from the final DiF model in Mplus and using these to create a distribution of factor scores, using 0.25sd bands across the range of underlying trait tested. These were then used alongside the expected scores of those same bands to create mean expected scores and standard deviations. The standard deviations and the sample size of the alternative group are used to generate an item pooled standard deviation, which is then used as the denominator of the calculation as shown below. The difference in the mean expected item scores, calculated using the parameters of the different groups is then divided by the item pooled standard deviation to provide the ESSD. This calculation allows us to classify the impact of DiF on the expected score of each item, according to Cohen's *d* classification, as either trivial ($\text{ESSD} < 0.2$), small ($0.2 \leq \text{ESSD} < 0.5$), moderate ($0.5 \leq \text{ESSD} < 0.8$) or large ($\text{ESSD} \geq 0.8$) (Cohen, 1988, Meade, 2010).

Item discrimination and difficulty parameters from the final DiF model were extracted and interpreted. Discrimination parameters examine how closely an item is related to the underlying QoL or wellbeing of respondents and were therefore used as a test of content and construct validity. Difficulty parameters assess over what levels of QoL the item is able to discriminate the trait level of respondents. Response distributions were also examined for floor and ceiling effects to further signal issues with discrimination. Item characteristic curves (ICCs) were used to examine the behaviour of the response categories. They also give a visual representation of the ability of the item to discriminate the underlying trait level of respondents, as judged by the height of the ICCs relative to the height of the ICCs of other items. Item information curves indicate over what range the item is best able to precisely discriminate and total test information curves provide information on the internal consistency reliability of the test at each level of latent trait. Total test information is analogous to calculating Cronbach's alpha at every level of underlying trait, with more information equivalent to higher Cronbach's alpha and higher internal consistency. Total test information ≥ 5 equals Cronbach's alpha ≥ 0.8 , a common cut-off for acceptable internal consistency in the literature (Fayers and Machin, 2016). Rates of missing data were also examined to look for issues with acceptability.

The stability of results was examined by taking the final DiF model obtained from the analyses in the development sample and rerunning this model in the validation sample. IRT parameters, expected scores for each age group and the ESSD estimates of the effect size of DIF were compared across the development and validation samples.

Several sensitivity analyses were undertaken to test results. First an alternative age cut-off of 75 was tested to examine whether the cut-off chosen (age 65) impacted results. Second, for the two shorter measures only single factor models were possible as two factor models would have been under-identified. However, some of the items within these measures may have been covering slightly different underlying concepts and therefore belong to other factors. When forced into a single factor these items may appear to perform poorly, as they do not relate well to the dominant factor measured by the instrument. However, if allowed into a factor which assesses the specific concept they are measuring, they may be seen to perform much better. As the EQ-5D-5L, ONS-4 and SF-12v2 measures are all found in the HIPO dataset a combined model, including these three instruments, was run to test whether the factors found in the individual measure models stuck and whether the item performance results hold.

4.3.6 Data preparation

To ease interpretation all items were coded so that higher numbered responses indicated better QoL and wellbeing. This involved reverse coding all EQ-5D and ASCOT items, items 1 and 8-10 from the SF-12v2 and the anxiety item from the ONS-4. MPlus only allows categorical variables to present 10 response options, which created a problem for the 11 response options of the ONS-4 questions. Prior to analysis the response options 0 and 1, signalling the lowest levels of wellbeing were merged for all respondents. This was done after the anxiety question was reverse coded so that the original categories 9 and 10 were combined for this item. These categories were chosen for merging due to the low response rates in these categories (evidenced in Appendix 9) and the fact that they all represent categories at the lowest end of the wellbeing scale.

This secondary data analysis was ethically approved by University of Sheffield Research Ethics Committee on 5th April 2017. The approval letter is shown in Appendix 10.

4.4 Results

4.4.1 Sample characteristics

Table 6 displays some basic characteristics of the three datasets, as a starting point from which to compare them. The full HIPO sample (EQ-5D-5L, SF-12v2, ONS-4) contained 6,351 patients recently discharged from the Cardiff and Vale NHS trust hospitals in 2014, of which, 42.8% were over the age of 65, 50.2% were female and 58.6% were married. The full HSE sample contained 7,870 adults (aged 18+) who received the self-completion booklet in which the WEMWBS appeared. The HSE sample tended to be younger than the HIPO, with 26.3% of respondents over 65 but a similar proportion of respondents were female (54.1%). A notably smaller proportion of the HSE were married and widowed, however there was a sizable 20% of respondents with no marital status recorded in this sample so these proportions may not be an accurate reflection. The self-reported general health of the HIPO sample tended to be slightly lower than the HSE. This may be because the HIPO sample is older and is made up of people recently discharged from hospital and therefore they might be more likely to be experiencing lingering health issues. The HIPO sample were also more likely to report receiving both formal and informal care than the HSE sample and a higher number of hours of care. The full ACSC (ASCOT) sample contains 69,081 respondents. It is most different of the three datasets, with all its respondents receiving formal care of some kind and most reporting to receive informal help, while the majority of the HIPO and HSE do not. This is reflected in the lower ratings of self-reported general health in the ASCS compared to the other surveys.

The HIPO model development sample contains 1,828 adults aged 18-64 and 1,348 aged over 65, while the ASCS contains 13,709 younger and 20,823 older adults and the HSE contains 2,719 respondents aged 18-64 and 940 over 65s. Appendices 11-13 show the sample characteristics of the development and validation samples for

each dataset. No significant differences were found between the development and validation datasets on any of the variables tested in Appendices 11-13. From this point on, the results presented will refer to the model development sample of each dataset, until the model validation results are presented at the end of the findings.

Table 6 - Sample characteristics (development and validation samples combined)

Characteristic	HIPO N = 6,351	HSE N = 7,253	ASCS N = 69,081
Age, n (%)			
18-64	3632 (57.2)	5378 (74.1)	27256 (39.5)
65+	2719 (42.8)	1875 (25.9)	41755 (60.5)
Gender, n (%)			
Female	3187 (50.2)	4067 (56.1)	41379 (60.0)
Marital Status, n (%)			
Married	3620 (57.0)	3905 (53.8)	N/A
Single	830 (13.1)	1249 (17.2)	
Divorced/separated	595 (9.4)	711 (9.8)	
Cohabiting	398 (6.3)	833 (11.5)	
Civil partnership	34 (0.5)	7 (0.1)	
Widowed	700 (11.0)	546 (7.5)	
General Health, n (%)			
Excellent/very good	2040 (32.1)	2334 (32.2)	9927 (14.4)
Good	1901 (29.9)	3093 (42.6)	18468 (26.8)
Fair	1590 (25.0)	1297 (17.9)	27097 (39.3)
Poor/bad/very bad	729 (11.5)	526 (7.3)	11916 (17.3)
Accommodation, n (%)			
Community	6025 (94.9)	7870 (100)	52158 (75.6)
Residential/Nursing home	77 (1.2)	0 (0)	16853 (24.4)
Informal Care			
Received, n (%)	1586 (25.0)	314 (4.3)	66227 (96.0)
Mean Hours last week*(SD)	31.0 (46.1)	18.5 (28.3)	N/A
Formal Care			
Received>0, n (%)	756 (11.9)	78 (1.1)	69018 (100)
Mean Hours last week* (SD)	11.6 (29.9)	8.76 (14.7)	N/A

*In those who received some care

4.4.2 Response distributions and missing data

EQ-5D-5L (HIPO validation dataset)

No floor effects were found in either age group (Table 7). However, there were substantial EQ-5D-5L ceiling effects in both age groups with large proportions of younger people reporting no problems in self-care (75.2%), mobility (58.8%), anxiety/depression (53.6%) and usual activities (47.5%) and the majority of older adults reporting no problems in self-care (66.2%) and anxiety (56.9%). A smaller proportion of older adults than young people responded no problems to all items except anxiety/depression. Missing data rates were low, with the maximum being 2.2% of older adults failing to provide a response to the anxiety/depression item.

SF-12v2 (HIPO validation dataset)

While no floor effects were seen, ceiling effects were also found for more than half of the SF-12v2 items, including the two physical role items (33.8% and 33.4%), two emotional role items (47.9% and 50.3%), pain (34.3%) and social activities (41.2%) in younger adults and the two emotional role questions (43.9% and 44.7%), downhearted/low (36.4%) and social activities (39.4%) in older adults (Table 8).

Missing data rates were mostly below 3% in younger adults, with the exception of 3.8% for the emotional role carefully question. However substantially more of the SF-12v2 questions had missing data rates above 3% in older adults including: stairs (4.0%), physical role limited (5.5%), both emotional role questions (4.5% and 9.1%), calm/peaceful (3.0%), energy (4.1%) and downhearted/low (3.4%). The particularly high missing response rate for the emotional role carefully item could signal an issue with this item in terms of acceptability or may indicate that it is easily missed in the layout of the questionnaire. It is interesting to note that for the three sets of items presented in pairs (moderate activities and stairs; the physical role items and the emotional role items), the rate of missing responses in the 2nd item in the pair was approximately double that of the 1st item. This could signal an issue with either layout or item redundancy within each pair.

Table 7 – EQ-5D-5L response distributions, n (%) by age group

18-64						EQ-5D Item	over 65						P-value*
Ext	Sev	Mod	Slight	None	Miss		Ext	Sev	Mod	Slight	None	Miss	
19 (1.0)	173 (9.4)	260 (14.2)	277 (15.2)	1075 (58.8)	24 (1.3)	Mobility	22 (1.6)	218 (16.2)	340 (25.2)	254 (18.8)	490 (36.4)	24 (1.8)	0.000
7 (0.4)	58 (3.2)	153 (8.4)	220 (12.0)	1374 (75.2)	16 (0.9)	Self-care	24 (1.8)	56 (4.2)	167 (12.4)	181 (13.4)	892 (66.2)	28 (2.1)	0.000
85 (4.7)	164 (9.0)	299 (16.4)	392 (21.4)	869 (47.5)	19 (1.0)	Usual activities	74 (5.5)	164 (12.2)	317 (23.5)	335 (24.9)	435 (32.3)	23 (1.7)	0.000
62 (3.4)	187 (10.2)	410 (22.4)	607 (33.2)	541 (29.6)	21 (1.2)	Pain/ discomfort	23 (1.7)	175 (13.0)	372 (27.6)	432 (32.1)	319 (23.7)	27 (2.0)	0.000
33 (1.8)	91 (5.0)	249 (13.6)	456 (25.0)	979 (53.6)	20 (1.1)	Anxiety/ depression	9 (0.7)	27 (2.0)	188 (14.0)	327 (24.3)	767 (56.9)	30 (2.2)	0.000

Sev=severe Mod=moderate Miss=missing Acts=activities dep=depression *Chi-square test

Table 8 – SF-12v2 response distributions, n (%) by age group

18-64						SF-12v2 Item	Over 65						P-value*
Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Miss		Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Miss	
214 (11.7)	392 (21.4)	509 (27.8)	520 (28.5)	172 (9.4)	21 (1.2)	1.General health	152 (11.3)	400 (29.7)	445 (33.0)	278 (20.6)	48 (3.6)	25 (1.9)	0.000
421 (23.0)	580 (31.8)	806 (44.1)	N/A	N/A	21 (1.2)	2.Moderate Activities	469 (34.8)	516 (38.3)	335 (24.9)	N/A	N/A	28 (2.1)	0.000
462 (25.3)	512 (28.0)	822 (45.0)	N/A	N/A	32 (1.8)	3.Stairs	544 (40.7)	440 (32.6)	310 (23.0)	N/A	N/A	54 (4.0)	0.000
225 (12.3)	288 (15.8)	330 (18.1)	332 (18.2)	618 (33.8)	35 (1.9)	4.Phys Role accomplish	190 (14.1)	272 (20.2)	373 (27.7)	235 (17.4)	241 (17.9)	37 (2.7)	0.000
256 (14.0)	269 (14.7)	298 (16.3)	346 (18.9)	611 (33.4)	48 (2.6)	5.Phys Role limited	198 (14.7)	259 (19.2)	351 (26.0)	227 (16.8)	239 (17.7)	74 (5.5)	0.000
125 (6.8)	182 (10.0)	282 (15.4)	337 (18.4)	876 (47.9)	26 (1.4)	6.Emot Role accomplish	73 (5.4)	167 (12.4)	240 (17.8)	215 (16.0)	592 (43.9)	61 (4.5)	0.000
113 (6.2)	164 (9.0)	265 (14.5)	298 (16.3)	919 (50.3)	69 (3.8)	7.Emot Role careful	76 (5.6)	119 (8.8)	227 (16.8)	201 (14.9)	602 (44.7)	123 (9.1)	0.000
154 (8.4)	293 (16.0)	258 (14.1)	468 (25.6)	627 (34.3)	28 (1.5)	8.Pain	86 (6.4)	271 (20.1)	233 (17.3)	337 (25.0)	384 (28.5)	37 (2.7)	0.000
120 (6.6)	285 (15.6)	468 (25.6)	773 (42.3)	157 (8.6)	25 (1.4)	9.Calm/ peaceful	38 (2.8)	151 (11.2)	308 (22.9)	606 (45.0)	205 (15.2)	40 (3.0)	0.000
281 (15.4)	369 (20.2)	486 (26.6)	549 (30.0)	113 (6.2)	30 (1.6)	10.Energy	244 (18.1)	284 (21.1)	388 (28.8)	326 (24.2)	51 (3.8)	55 (4.1)	0.000
86 (4.7)	192 (10.5)	446 (24.4)	538 (29.4)	537 (29.4)	29 (1.6)	11.Down/ low	39 (2.9)	86 (6.4)	284 (21.1)	403 (29.9)	490 (36.4)	46 (3.4)	0.000
182 (10.0)	220 (12.0)	350 (19.2)	304 (16.6)	753 (41.2)	19 (1.0)	12.Social activities	130 (9.6)	191 (14.2)	269 (20.0)	195 (14.5)	531 (39.4)	32 (2.4)	0.012

Items 1 and 8-10 were reverse coded so that category 1 represents the lowest level of health for each item. *Chi-square test

Cat=category Miss=missing Mod=moderate Phys=physical Emot=emotional Down=downhearted

ASCOT (ASCS validation dataset)

While no floor effects were found in either age group, substantial ceiling effects were observed in all ASCOT items except control for under 65s and in all except control and occupation in older adults (Table 9). Missing data rates were above 3% for dignity in under 65s and for social participation (3.0%), occupation (3.7%) and dignity (4.7%) in over 65s.

Table 9 – ASCOT response distributions, n (%) by age group

18-64					ASCOT	Over 65					P-value*
Cat 1	Cat 2	Cat 3	Cat 4	Miss		Cat 1	Cat 2	Cat 3	Cat 4	Miss	
471 (3.4)	2079 (15.2)	6111 (44.6)	4813 (35.1)	235 (1.7)	1.Control	1325 (6.4)	4020 (19.3)	8892 (42.7)	6195 (29.8)	391 (1.9)	0.000
153 (1.12)	706 (5.2)	3969 (29.0)	8649 (63.1)	232 (1.7)	2.Personal clean	119 (0.6)	833 (4.0)	8589 (41.3)	10788 (51.8)	494 (2.4)	0.000
194 (1.4)	602 (4.4)	3749 (27.4)	8853 (64.6)	311 (2.3)	3.Food/ Drink	159 (0.8)	921 (4.4)	6570 (31.6)	12570 (60.4)	603 (2.9)	0.000
366 (2.7)	640 (4.7)	3443 (25.1)	8998 (65.6)	262 (1.9)	4.Safety	256 (1.2)	718 (3.5)	5458 (26.2)	13861 (66.6)	530 (2.6)	0.000
871 (6.4)	2158 (15.7)	3890 (28.4)	6474 (47.2)	316 (2.3)	5.Social part	867 (4.2)	3561 (17.1)	7398 (35.5)	8366 (40.2)	631 (3.0)	0.000
569 (4.2)	2834 (20.7)	3917 (28.6)	6125 (44.7)	264 (1.9)	6.Occupation	1870 (9.0)	5408 (26.0)	6946 (33.4)	5840 (28.1)	759 (3.7)	0.000
161 (1.2)	655 (4.8)	3564 (26.0)	9009 (65.7)	320 (2.3)	7.Accommodation	60 (0.3)	672 (3.2)	6412 (30.8)	13111 (63.0)	568 (2.7)	0.000
205 (1.5)	966 (7.1)	3090 (22.5)	8969 (65.4)	479 (3.5)	8.Dignity	192 (0.9)	1553 (7.5)	6617 (31.8)	11482 (55.1)	979 (4.7)	0.000

All items were reverse coded so that category 1 represents the lowest level of SCRQoL for each item. Cat=category Miss=missing Clean=cleanliness and comfort Part=participation

*Chi-square test

WEMWBS (HSE dataset)

In younger adults only WEMWBS feeling loved (36.4%) exhibited a ceiling effect while, in older adults, ceilings were observed for feeling loved (38.6%), thinking clearly (30.4%) and able to make up own mind (37.9%) (Table 10). Interestingly a higher proportion of older adults responded “all of the time” in all except two items (optimistic about the future and energy to spare). This could be anticipated as these are areas that older adults may be expected to struggle more with than younger adults. Missing data rates were above 3% in all WEMWBS items in both groups.

Table 10 – WEMWBS response distributions, n (%) by age group

18-64						WEMWBS Item	over 65					P-value*	
None	Rare	Some	Often	All	Miss		None	Rare	Some	Often	All		
113 (4.2)	288 (10.6)	952 (35.1)	945 (34.8)	293 (10.78)	128 (4.7)	1.Optimistic about future	64 (6.8)	135 (14.4)	367 (39.0)	240 (25.5)	88 (9.4)	46 (4.9)	0.000
82 (3.0)	171 (6.3)	824 (30.3)	1138 (41.9)	379 (13.9)	125 (4.6)	2.Useful	32 (3.4)	63 (6.7)	337 (35.9)	317 (33.7)	143 (15.2)	48 (5.1)	0.001
66 (2.4)	358 (13.2)	1085 (39.9)	893 (32.8)	198 (7.3)	119 (4.4)	3.Relaxed	19 (2.0)	50 (5.3)	348 (37.0)	352 (37.5)	125 (13.3)	46 (4.9)	0.000
76 (2.8)	219 (8.1)	828 (30.5)	1113 (40.9)	361 (13.3)	122 (4.5)	4.Interested other people	15 (1.6)	50 (5.3)	298 (31.7)	350 (37.2)	179 (19.0)	48 (5.1)	0.000
142 (5.2)	628 (23.1)	1116 (23.1)	579 (21.3)	133 (4.9)	121 (4.5)	5.Energy to spare	89 (9.5)	201 (21.4)	399 (42.5)	171 (18.2)	33 (3.5)	47 (5.0)	0.000
41 (1.5)	138 (5.1)	837 (30.8)	1229 (45.2)	357 (13.1)	117 (4.3)	6.Deal problems	14 (1.5)	33 (3.5)	294 (31.3)	364 (38.7)	188 (20.0)	47 (5.0)	0.000
27 (1.0)	116 (4.3)	628 (23.1)	1287 (47.3)	543 (20.0)	118 (4.3)	7.Think clearly	11 (1.2)	25 (2.7)	217 (23.1)	351 (37.3)	286 (30.4)	50 (5.3)	0.000
41 (1.5)	219 (8.1)	949 (34.9)	1088 (40.0)	331 (12.2)	91 (3.4)	8.Feel good about self	18 (1.9)	65 (6.9)	351 (37.3)	311 (33.1)	164 (17.5)	31 (3.3)	0.000
45 (1.7)	194 (7.1)	829 (30.5)	1126 (41.4)	431 (15.9)	94 (3.5)	9.Close other people	11 (1.2)	57 (6.1)	303 (32.2)	367 (39.0)	170 (18.1)	32 (3.4)	0.305
48 (1.8)	215 (7.9)	899 (33.0)	1102 (40.5)	361 (13.3)	94 (3.5)	10.Confident	7 (0.7)	60 (6.4)	339 (36.1)	341 (36.3)	164 (17.5)	29 (3.1)	0.001
14 (0.5)	94 (3.5)	484 (17.8)	1247 (45.9)	788 (29.0)	92 (3.4)	11.Make up mind	8 (0.9)	17 (1.8)	174 (18.5)	355 (37.8)	356 (37.9)	30 (3.2)	0.000
40 (1.5)	137 (5.0)	519 (19.1)	941 (34.6)	990 (36.4)	92 (3.4)	12.Loved	13 (1.4)	41 (4.4)	196 (20.9)	295 (31.4)	363 (38.6)	32 (3.4)	0.431
65 (2.4)	233 (8.6)	802 (29.5)	1066 (39.2)	463 (17.0)	90 (3.3)	13.Interested new things	25 (2.7)	91 (9.7)	318 (33.8)	305 (32.5)	171 (18.2)	30 (3.2)	0.012
31 (1.1)	158 (5.8)	823 (30.3)	1258 (46.3)	357 (13.1)	92 (3.4)	14.Cheerful	6 (0.6)	51 (5.4)	263 (28.0)	419 (44.6)	172 (18.3)	29 (3.1)	0.005

Rare=rarely Miss=missing *Chi-square test

ONS-4 (HIPO dataset)

No floor effects were found but ceilings were found for worthwhile in over 65s (20.1%) and anxiety in both age groups (under 65s 35.8% and over 65s 41.8%) (Table 11). Interestingly over 65s were more likely than younger adults to respond in the top category, representing the highest level of wellbeing, for all items. Missing response rates were below 2% for all items in under 65s and 2-3% in over 65s, with the highest being 3.0% for worthwhile in over 65s.

Table 11 – ONS-4 response distributions, n (%) by age group

18-64				Response	Over 65			
Life Sat	Worth	Happy	Anxiety		Life Sat	Worth	Happy	Anxiety
103 (5.6)	75 (4.1)	91 (5.0)	85 (4.7)	Cat 1	45 (3.3)	39 (2.9)	29 (2.2)	26 (1.9)
79 (4.3)	71 (3.9)	69 (3.8)	79 (4.3)	Cat 2	36 (2.7)	39 (2.9)	37 (2.7)	49 (3.6)
111 (6.1)	91 (5.0)	77 (4.2)	100 (5.5)	Cat 3	64 (4.8)	46 (3.4)	52 (3.9)	67 (5.0)
95 (5.2)	84 (4.6)	107 (5.9)	88 (4.8)	Cat 4	80 (5.9)	57 (4.2)	61 (4.5)	59 (4.4)
180 (9.9)	150 (8.2)	156 (8.5)	152 (8.3)	Cat 5	179 (13.3)	120 (8.9)	122 (9.1)	126 (9.4)
173 (9.5)	123 (6.7)	142 (7.8)	110 (6.0)	Cat 6	93 (6.9)	89 (6.6)	79 (5.9)	67 (5.0)
265 (14.5)	231 (12.6)	263 (14.4)	142 (7.8)	Cat 7	187 (13.9)	133 (9.9)	178 (13.2)	91 (6.8)
398 (21.5)	377 (20.6)	366 (20.0)	194 (106)	Cat 8	282 (20.9)	288 (21.4)	259 (19.2)	125 (9.3)
250 (13.7)	319 (17.5)	328 (17.9)	203 (11.1)	Cat 9	196 (14.5)	225 (16.7)	237 (17.6)	144 (10.7)
158 (8.6)	277 (15.2)	206 (11.3)	654 (35.8)	Cat 10	155 (11.50)	271 (20.1)	262 (19.4)	564 (41.8)
21 (1.2)	30 (1.6)	23 (1.3)	21 (1.2)	missing	31 (2.3)	41 (3.0)	32 (2.4)	30 (2.2)
0.000	0.000	0.000	0.000	P-value*	0.000	0.000	0.000	0.000

Anxiety was reverse coded so that category 1 represents the lowest level of wellbeing

Sat=satisfaction Worth=worthwhile Cat=category *Chi-square test

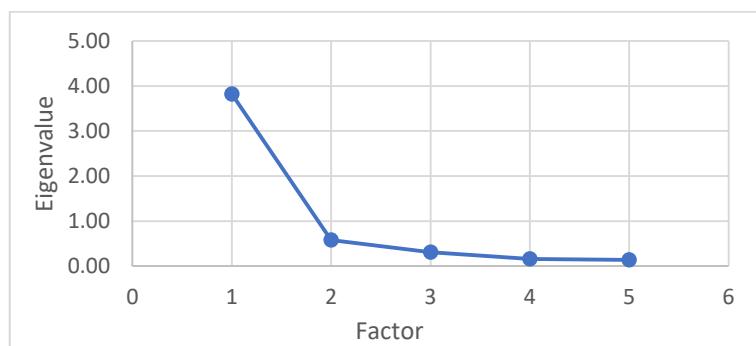
The differences in response distributions between age groups are interesting, with younger adults more likely to report no problems in questions relating to physical problems or occupation and older adults more likely to report no problems to mental health and wellbeing questions. This may suggest either that older adults have fewer issues in mental health or are less likely to recognise them as issues.

4.4.3 IRT model comparisons and assumption checking

EQ-5D-5L

The Geomin rotated EFA model had eigenvalues on the first and second factors of 3.8 and 0.58, which suggests a one-factor model. The loadings of the two-factor EFA suggested that only the anxiety/depression item would load onto the second factor. The scree plot also suggested a single factor model (Figure 6). Therefore, a single factor model was used. The model was first run as a two-parameter GRM using full information ML estimation. Equality of discrimination parameters between items was then tested by running a constrained one-parameter GRM and comparing nested model fit using the difference in -2^*LL test. This showed that the two-parameter GRM fit better (Table 12) and equality of item discriminations was rejected. Next a nominal model was run to check if the levels were being treated as ordinal, in the intended rank order. Model fit was compared to the GRM using the AIC and BIC statistics. The GRM was found to fit better than the nominal, suggesting no issues with category ordering. Finally, the model was run as a GPCM to see if this model demonstrated superior fit. The two-parameter GRM and GPCM were compared in terms of relative model fit using the AIC and BIC (Table 12), which confirmed that the two-parameter GRM fit better. The two-parameter GRM was therefore taken forward to the multiple-group DiF analysis phase.

Figure 6 – EQ-5D-5L scree plot

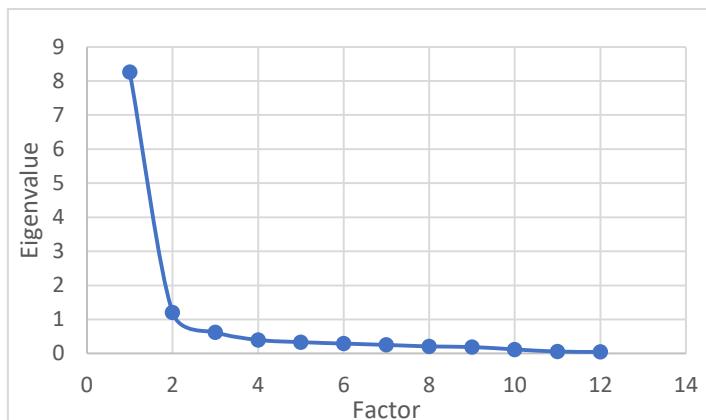


The MIs, which specify areas of local misfit in the model and potential improvements to model fit, did not suggest any issues with local dependence. The Mplus input files for each EQ-5D model tested are provided in Appendix 14. The code for the models tested for the other measures follow the same stages and are available on request. The code for the final DIF model of each of the remaining measures is provided as an example.

SF-12v2

The Geomin rotated EFA model had eigenvalues on the first, second and third factors of 8.3, 1.2 and 0.6, suggesting a two-factor model with items; general health, moderate activities, stairs, physical role accomplish, physical role limited, pain, energy and social activities grouped together in a physical health factor and items; emotional role accomplish, emotional role carefully, calm/peaceful, and downhearted/low in a mental health factor. The scree plot, shown in Figure 7, also suggested a two-factor solution. The split of the items between the two-factors made theoretical sense but it should be noted that this split is inconsistent with the split of items between the physical and mental component summary scales of the original scoring system of the SF-12v2 where items 1-5 and 8 made up the physical and 6, 7 and 9-12 made up the mental component. However, the split found in this dataset corresponds with previous SF-12v2 factor analysis findings in older groups (Cernin, Cresci et al., 2010, Resnick and Nahm, 2001). The single and two-factor model were both run in ML as two-parameter GRMs and compared in terms of fit. The AIC and BIC were lower for the two-factor model suggesting this fit better (Two-factor model AIC=75987.5 BIC=76332.7 One-factor model AIC=79017.8 BIC=79356.9). A bifactor model was run to test if there was sufficient unidimensionality for a single factor IRT. The global QoL factor explained 48.4% of the common variance while the physical and mental health factors explained 40.6% and 11.0%. The global factor did not explain substantially more of the common variance than the two separate factors, and failed to meet the minimum suggested cut-off of 50% of explained common variance (Reise, Bonifay et al., 2013), so sufficient unidimensionality was rejected and the two-factor model was taken forward.

Figure 7 – SF 12 Scree Plot



The resulting two-factor model was initially run as a two-parameter GRM and equality of discriminations was tested by comparing nested fit with a constrained one-parameter GRM with the same factor structure in which all item discriminations were forced to be equal. The difference in -2^*LL test again confirmed that the two-parameter GRM fit better (Table 12). This was then compared to a nominal model using AIC and BIC. Again, the nominal model had a poorer fit (Table 12), showing there were no issues with the rank order understanding of the item responses. Lastly, a GPCM model was tested and compared to the GRM using the AIC and BIC. These statistics showed that the two-parameter GRM fit best and this model was therefore taken forward to the multiple-group DiF analysis phase.

The MIs suggested some issues with local dependence, with large MIs for error covariances between the moderate activities and stairs items and the physical role accomplishment and physical role limited items. This may suggest item redundancy within these pairs as they cover very similar topics and are grouped in pairs. When the physical role limited item was removed, there was a substantial reduction on the discrimination parameter of the physical role accomplishment item and a change in rank from highest physical discrimination to second highest in over 65s, with the moderate activities becoming the highest. When the stairs item was removed there was little change in the discrimination of the moderate activities item and no change in ranking. This suggests there may be local dependence between the pair of physical role items. The discrimination parameters for the pair of physical role items were substantially higher than all other items in the physical health factor, while the discriminations of the emotional role items dominated in the remaining items in the mental health factor, again suggesting local dependence within these item pairs.

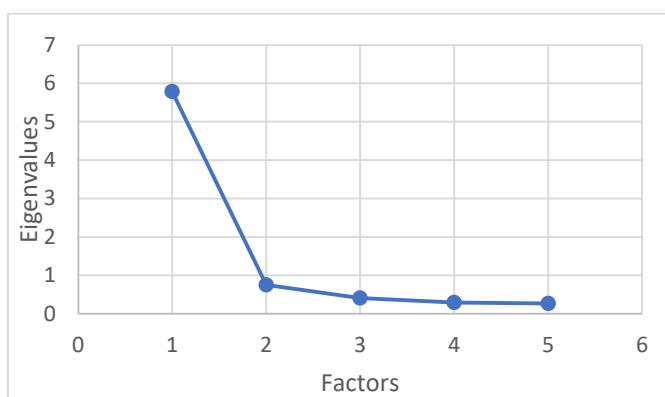
The suggestion of local dependence within these three pairs of items follows the conceptual structure of the SF-12v2, as these three pairs of items form three of the four multi-item scales within the measure (moderate activities and stairs = physical functioning scale; physical role accomplish and physical role limited = physical role scale; emotional role accomplish and emotional role carefully = emotional role scale; and calm/peaceful and downhearted/low = mental health scale). Methods have been proposed to deal with local dependence stemming from multi-item scales within a measure by forming testlets (Steinberg and Thissen, 1996). These methods are referred to as testlet analysis (TLA) from hereon. In the TLA approach, items which make up multi-item scales can be combined to form super items (or testlets), by

summing the scores of each respondent across these items. These super items are then treated as individual items in the IRT analyses. By treating each pair of items as an individual super item the issue of local independence can be solved while retaining the items for analysis and maintaining consistency with the conceptual design of the SF-12v2. Therefore this approach was undertaken in the following section, entitled SF-12v2 TLA.

SF-12v2 TLA

The eigenvalues on the first, second and third factors of the Geomin rotated EFA model were 5.79, 0.75 and 0.41, suggesting a single factor model. The scree plot (Figure 8) suggested either a single or two factor model. The split of items across the two-factor model was consistent with the split seen in the non-TLA analysis, with general health, pain, energy, social functioning and the physical functioning and physical role super items forming a physical health factor and the emotional role and mental health super items forming a mental health factor. The single and two-factor model were both run as two-parameter GRMs in ML and compared in terms of fit. The AIC and BIC were lower for the two-factor model suggesting this fit better (Two-factor model AIC=130108 BIC=130466 One-factor model AIC=130898 BIC=131249). A two-factor model was taken forward as this was consistent with the conceptual model of the SF-12v2 and provided superior fit.

Figure 8 – Scree plot SF-12v2 TLA



The two-factor model was initially run as a two-parameter GRM. Equality of item discriminations was tested by comparing nested fit with a constrained one-parameter GRM with the same factor structure. The difference in $-2 \times LL$ test confirmed that the two-parameter GRM fit better (Table 12). This model was then compared to a nominal

model to check whether there were issues in the use or understanding of the ordinal categories. The AIC and BIC showed that the nominal model fit worse. The two-parameter GRM was then compared to a two-parameter GPCM using AIC and BIC. Again, the GPCM had a poorer fit (Table 12), and therefore the two-parameter GRM was taken forward to the multiple-group DiF analysis phase.

ASCOT

The EFA eigenvalues on the first, second and third factors were 4.4, 0.7 and 0.66. This and the scree plot suggested a single factor model. The split of the items in the two-factor made theoretical sense with items; control, social participation and occupation, grouping into an external activities factor and items; personal cleanliness/comfort, food/drink, safety, accommodation cleanliness/comfort and dignity grouping into a stable environment factor. The two-factor model fit better in terms of AIC and BIC (Two-factor model AIC= 459206.8 BIC= 459485.5. One-factor model AIC= 460660.1 BIC= 460930.3). Sufficient unidimensionality was then tested in a bifactor model. The global SCRQoL factor explained the majority of the common variance (78%), while the environment and activities factors explained 12% and 10% respectively. The global factor surpassed the recommended cut-off for explained common variance of 75% (Reise, Bonifay et al., 2013). It was therefore considered that the global factor dominated the individual factors enough to claim sufficient unidimensionality and the single factor model was taken forward.

Again, the two-parameter GRM was shown to fit significantly better than the constrained one-parameter GRM by the difference in $-2 \times LL$ nested model fit test and equality of discrimination parameters across items was rejected (Table 12). The two-parameter nominal model was also shown by the AIC and BIC to have a worse fit than the GRM suggesting that there were no issues with the use of the ordering of response categories and again the two-parameter GRM model taken forward for DiF analysis

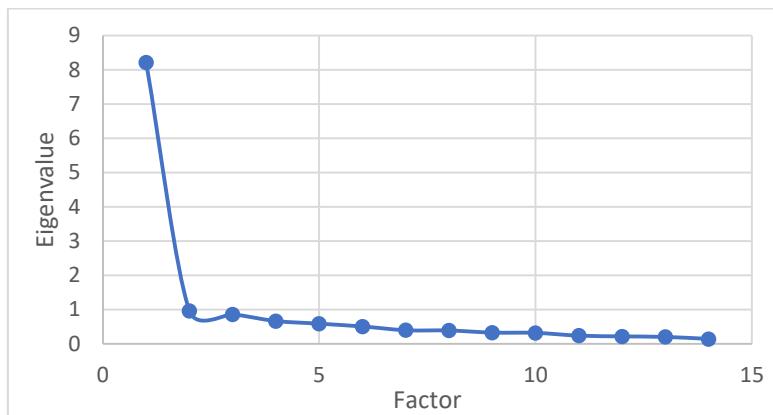
Large MIs for error covariances were found between the social and occupation items and the personal and accommodation cleanliness/comfort items. Again, this is not surprising as the wording and topic of these item pairs are similar. However, when one item from each pair was removed and the model rerun there were no substantial effects on the unstandardized discrimination parameter of the remaining item or its

discrimination parameter ranking. Again, all items were retained to get full information about item performance.

WEMWBS

The Geomin rotated EFA model had eigenvalues on the first, second and third factors of 8.2, 0.96 and 0.9 respectively. The scree plot (Figure 9) suggested a one-factor model. The split of the items in the two-factor model made theoretical sense. Feeling optimistic, useful, interested in other people, close to other people, loved and interested in new things grouped together in the first factor. These items all seemed to be considered wellbeing in relation to external concepts, people and events and is therefore referred to as external wellbeing. The other items were more internal concepts such as feeling relaxed, confident, cheerful and good about oneself, with energy to spare, dealing with problems well, thinking clearly and able to make up your own mind.

Figure 9 WEMWBS Scree Plot



Fit of the single and two-factor GRM ML models were compared. The AIC and BIC were lower for the two-factor model suggesting this fit better (Two-factor model AIC=101249.5 BIC=101689.6 One-factor model AIC=101710.0 BIC=102143.9). A bifactor model was then run to test sufficient unidimensionality for a single factor IRT. The global wellbeing factor explained 49.9% of the common variance and the external and internal factors explained 23.8% and 26.4% respectively. The common variance explained by the global factor failed to meet the minimum suggested cut-off of 50% (Reise, Bonifay et al., 2013). The non-global factors contributed substantially to explaining common variance and therefore sufficient unidimensionality was rejected and the two-factor model was taken forward.

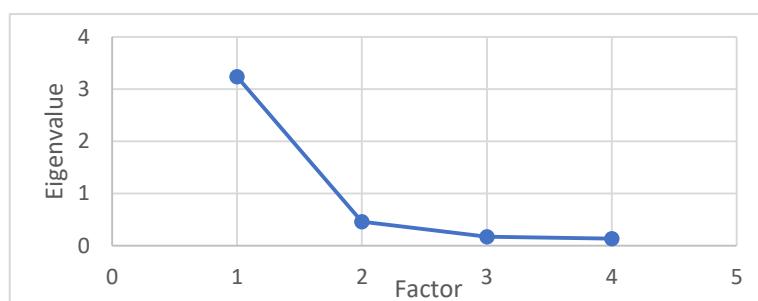
Again, the resulting two-factor model was initially run as a two-parameter GRM in ML and equality of discriminations was tested by comparing nested fit with a constrained one-parameter GRM with the same factor structure. The difference in -2^*LL test again confirmed that the two-parameter GRM fit significantly better (Table 12). This was then compared, first to a nominal model and then to a two-parameter GPCM using AIC and BIC. Again, the nominal model fit worse, showing there were no issues with the rank order understanding of the item responses (Table 12). The GPCM also fit worse and so the two-parameter GRM was therefore taken forward to the multiple-group DiF analysis phase.

Local dependence was suggested by the MIs between dealing with problems and thinking clearly; feeling close to other people and loved; and feeling good about oneself and confident. This is likely due to redundancy within these pairs as they cover very similar concepts. However, when one item from each pair was removed and the model rerun there were no substantial effects on the unstandardized discrimination parameter of the remaining item or its discrimination parameter ranking. All items were left in the model as the LD test of removing one item from the pair did not show any big impact. Again, no items were retained to gain the maximum amount of information about the performance of the measure.

ONS-4

The Geomin rotated EFA model had eigenvalues on the first and second factors of 3.2 and 0.46, which suggests a one-factor model. The scree plot also suggested a single factor model (Figure 10). Therefore, a single factor model was adopted.

Figure 10 – ONS-4 scree plot



The model was first run as a two-parameter GRM. Equivalence of discrimination parameters between items was then tested by running a constrained one-parameter

GRM and comparing nested model fit using the difference in -2*LL test. This showed that the two-parameter GRM fit better and the assumption of equal discriminations was rejected (Table 12). Next the model was run as a two-parameter nominal categories model to check if the levels were being treated as ordinal, in the intended rank order, and finally a GPCM was tested. The GRM, nominal model and GPCM were compared in terms of nested model fit using the AIC and BIC, which confirmed that the GRM fit better and there were no issues with response categories being used out of order. The two-parameter GRM was therefore taken forward to the multiple-group DiF analysis phase. The MIs did not suggest any issues with local dependence.

Table 12 – Model comparison nested and relative fit statistics

Model Type	Nested Model Fit					Relative Model Fit	
	-2*LL	df	Diff -2*LL	Diff df	p value	AIC	BIC
EQ-5D-5L							
Constrain-GRM	31209.0	21	885.4	4	<0.001	31251.0	31378.2
GRM	30323.6	25				30373.6	30525.0
Nominal						30594.3	30830.4
GPCM						30526.8	30678.2
SF-12v2							
Constrain-GRM	78582.3	47	2708.8	10	<0.001	78676.3	78960.9
GRM	75873.5	57				75987.5	76332.7
Nominal						79480.1	80007.0
GPCM						76726.8	77072.0
SF-12v2 TLA							
Constrain-GRM	130878.1	47	876.2	6	<0.001	130972.1	131289.3
GRM	130001.9	53				130107.9	130465.6
Nominal						134742.6	135329.7
GPCM						131059.3	131376.5
ASCOT							
Constrain-GRM	462381.7	25	1785.6	7	<0.001	462431.7	462642.8
GRM	460596.1	32				460660.1	460930.6
Nominal						462978.1	463374.9
GPCM						461554.8	461825.0
WEMWBS							
Constrain-GRM	102234.7	59	1127.2	12	<0.001	102352.7	102718.4
GRM	101107.5	71				101249.5	101689.6
Nominal						102582.4	103270.5
GPCM						102814.9	103256.0
ONS-4							
Constrain-GRM	43747.0	37	1011.6	3	<0.001	43821.0	44044.9
GRM	42735.4	40				42815.4	43057.5
Nominal						45171.5	45601.2
GPCM						43368.5	43610.5

4.4.4 DiF model parameters and DiF impact

Factor structures were found to remain the same across age groups for all measures. Parameters shown in this section are those which result from the final model of the DiF process. The order of models tested, the parameters freed at each stage and the fit of each DiF model tested in the process are shown in the DiF model fit tables in Appendix 15. The discrimination and difficulty parameters found to differ across age groups and exhibit DiF are shown in bold in Tables 13-22.

EQ-5D-5L

The absolute fit statistics for the multiple-group two-parameter GRM IRT model were good. The CFI was 0.999 (good \geq 0.96). Mean (95% CI) RMSEA was 0.047 (0.037, 0.058). The density plots of HRQoL in Figure 11 show the distribution of the estimated levels of HRQoL of under and over 65s respectively. The EQ-5D-5L estimates the HRQoL of under 65s to be between approximately 3.25 SDs below and 0.75 SDs above the mean level of HRQoL. In over 65s the EQ-5D-5L predicts the HRQoL of respondents to be between 3.75 SDs below and 0.5 SDs above mean HRQoL. This suggests both distributions are negatively skewed, particularly over 65s. The spike at the top end of the distribution in both age groups signals the ceiling of the measure, where a substantial proportion of respondents have hit the highest level of HRQoL which the instrument is capable of measuring. Therefore, the EQ-5D-5L is not able to discriminate the HRQoL of respondents above 0.75 SDs above the mean HRQoL level.

Unstandardised discrimination parameters (Table 13) ranged from 2.64 for mobility to 0.86 for anxiety/depression for both age groups, suggesting mobility is most closely related to HRQoL in both age groups and anxiety/depression least related. However, anxiety/depression is still relevant to HRQoL with a standardised discrimination parameter of 0.634 in over 65s. Non-uniform DiF was indicated by higher discriminations for pain/discomfort and self-care in under 65s. This suggests that these concepts are less closely related to the HRQoL of over 65s than under 65s while mobility, usual activities and anxiety/depression are equally well related to HRQoL regardless of age. This could be anticipated as over 65s may be more used to issues in self-care and pain/discomfort and may have adapted.

Figure 11 – Density plots of predicted HRQoL from the EQ-5D-5L

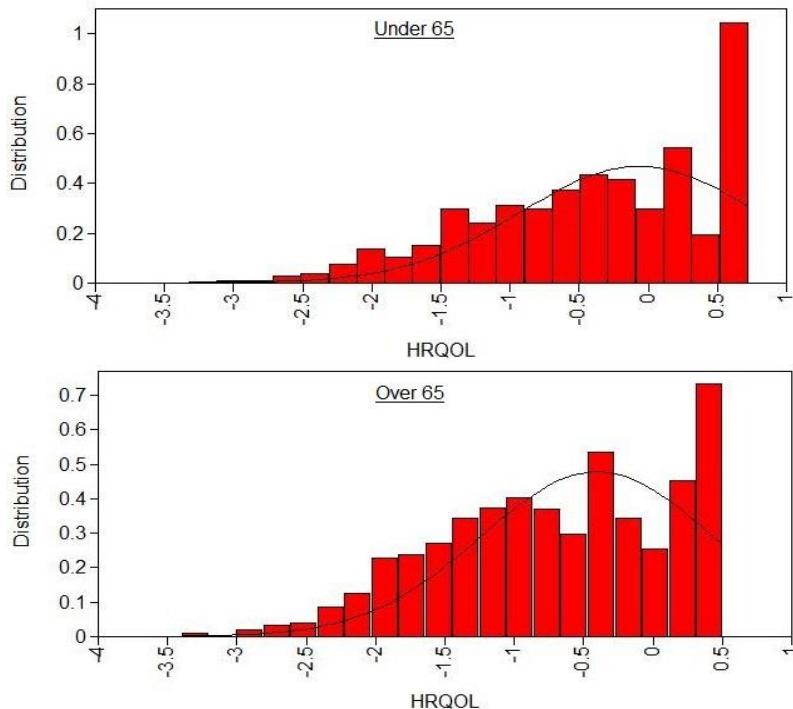


Table 13 – EQ-5D-5L factor structures, unstandardised discrimination parameters and absolute model fit statistics

	Unstandardised discrimination parameters (SEs)	
	Under 65s	Over 65s
EQ-5D	HRQoL	HRQoL
1.Mobility	2.64 (0.12)	2.64 (0.12)
2.Self-care	2.47 (0.13)	1.86 (0.10)
3.Usual Activities	2.32 (0.09)	2.32 (0.09)
4.Pain/discomfort	1.69 (0.05)	1.30 (0.06)
5.Anxiety/depression	0.86 (0.03)	0.86 (0.03)
Factor Mean	0	-0.359 (0.04)
Factor Variance	1	0.914 (0.07)
Model Fit	RMSEA mean (90% CI)	CFI
	0.047 (0.037, 0.058)	0.999

Discrimination parameters in bold exhibit non-uniform DiF as they were found to differ between under and over 65s

The difficulty parameters (b1-b4 Table 14), represent the amount of HRQoL required to have a 50% probability of responding above a certain category, signalling higher HRQoL. Anyone with above average HRQoL is most likely to respond “no problems” for self-care, anxiety/depression and mobility as these items all have negative b4 parameters. These low b4s correspond with the substantial ceiling effects found for

the EQ-5D-5L. Uniform DiF was indicated by difficulty parameters which differed between age groups. All difficulty parameters for pain/discomfort and anxiety/depression were lower in over 65s, meaning over 65s require less HRQoL to be more likely to respond in the next category up than younger adults and therefore, older adults are more likely to respond higher to pain/discomfort and anxiety/depression than a younger person with the same level of HRQoL. Among the remaining difficulty parameters exhibiting DiF, older adults tended to be more likely to respond higher for self-care and usual activities but lower to mobility than a younger adult with the same HRQoL level.

Table 14 – EQ-5D-5L difficulty parameters

Difficulty Parameters (SEs)									
Under 65s				EQ-5D	Over 65s				
b1	b2	b3	b4		b1	b2	b3	b4	
-2.51 (0.21)	-1.31 (0.15)	-0.72 (0.12)	-0.26 (0.09)	1.Mobility	-2.51 (0.21)	-1.31 (0.15)	-0.52 (0.11)	-0.02 (0.11)	
-2.87 (0.35)	-1.94 (0.21)	-1.27 (0.16)	-0.76 (0.13)	2.Self-care	-2.66 (0.17)	-2.06 (0.16)	-1.33 (0.13)	-0.86 (0.11)	
-1.82 (0.13)	-1.19 (0.11)	-0.57 (0.09)	0.08 (0.07)	3.Usual Activities	-2.03 (0.16)	-1.32 (0.13)	-0.57 (0.09)	0.08 (0.07)	
-2.18 (0.10)	-1.26 (0.06)	-0.42 (0.05)	0.61 (0.06)	4.Pain/discomfort	-2.83 (0.10)	-1.64 (0.06)	-0.55 (0.05)	0.50 (0.07)	
-3.21 (0.09)	-2.28 (0.06)	-1.26 (0.05)	-0.16 (0.04)	5.Anxiety/depression	-4.08 (0.16)	-3.26 (0.10)	-1.80 (0.06)	-0.67 (0.05)	

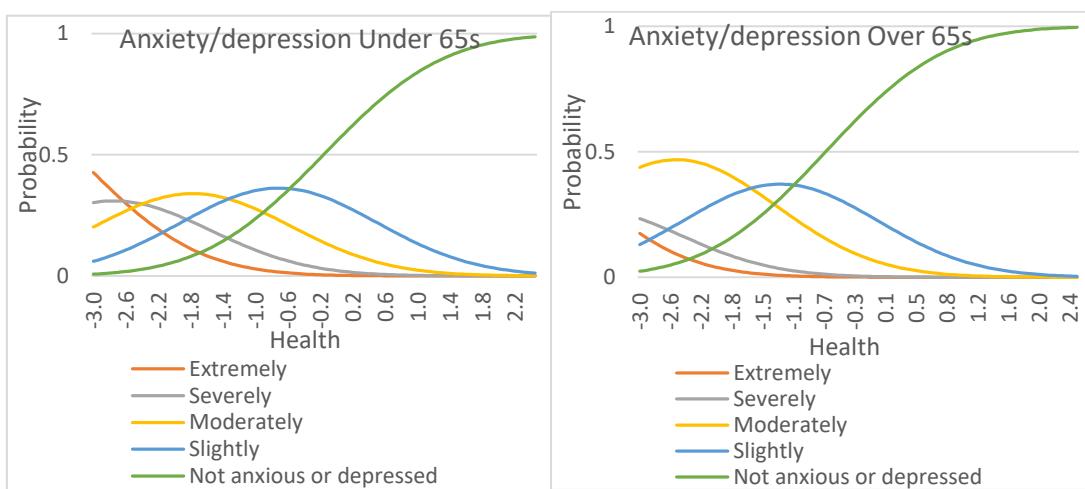
Where b1 represents the amount of QoL required to have a 50% probability of responding in the category signalling the lowest level of QoL and 50% chance of responding higher. Difficulty parameters in bold exhibit uniform DiF as they were found to differ between under and over 65s

Constraining the unstandardized residual variances of both groups to 1 did not significantly impact the fit of the model ($p\text{-value}=0.72$) indicating that the amount of item variance not accounted for by the factor was the same across groups. The factor mean differed between the two age groups, as shown in Table 13. The mean level of health is 0.359 SDs lower in the over 65 group. This difference was significant ($P\text{-value}<0.000$). The factor variance was found to differ between groups. As shown in Table 13, the over 65 group were slightly less variable in health with a factor variance of 0.914, compared to the constrained factor variance of 1 in under 65s. Factor invariance was then tested to examine whether constraining the factor variance to

equal 1 in both groups significantly impacted the model. Again, the DIFFTEST was insignificant ($p\text{-value}=0.248$), suggesting that constraining the factor variance to 1 in both groups did not significantly impact model fit and group factor variances do not differ significantly.

The ICCs for older and younger adults for each item are shown in Appendix 16. Item levels mostly behaved as expected, with distinct ordered categories each being the most likely over a range of HRQoL. However, there were issues with the ICCs for the anxiety/depression item. For anxiety/depression in over 65s (Figure 12) “extremely” and “severely” are never the most likely response options over the range of HRQoL considered (-3 to +2.8 SDs about mean HRQoL), while they do become the most likely within this range in under 65s. We would expect very few respondents to have a HRQoL level below 3sd below mean HRQoL. This suggests that over 65s are less likely to use these categories than under 65s and that they will be used very infrequently for the anxiety/depression item.

Figure 12 – Examples of problematic Item characteristic curves from the EQ-5D-5L



Items mobility, self-care and usual activities provide the highest levels of information (Figure 13) for those with below average HRQoL. Pain/discomfort and anxiety/depression provide lower levels of information. Total information (Figure 14) is above 5 (Cronbach's alpha ≥ 0.8) between -3 SDs and +0.75 SDs about the mean for younger adults and +0.7 SDs in older adults, indicating good internal reliability and precision of measurement within this range. The EQ-5D-5L does not have good internal reliability to predict the HRQoL in those above 0.75 SDs above mean HRQoL, due to the ceiling effect.

Figure 13 – EQ-5D-5L Item Information by age group

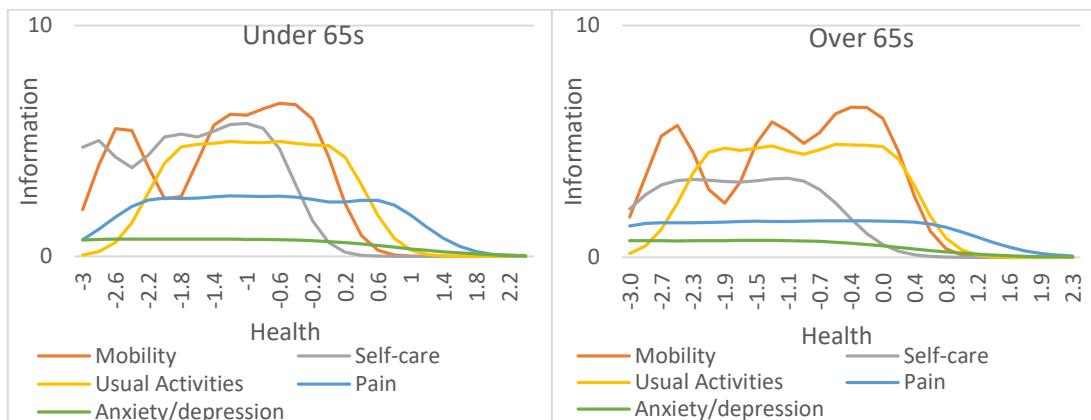


Figure 14 – EQ-5D-5L Total Information

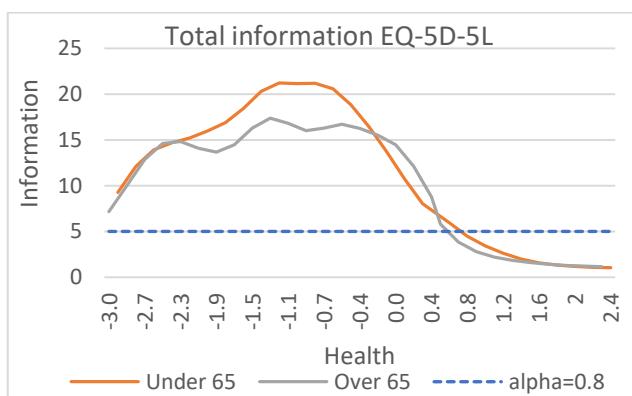


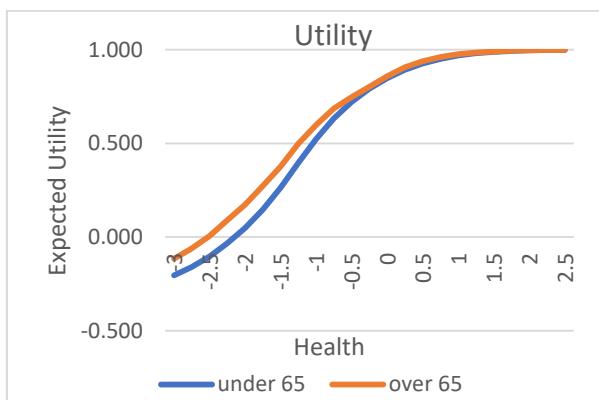
Table 15 – EQ-5D-5L Item level and total measure DiF effect size

Item	Mobility	Self-care	Usual Activities	Pain/discomfort	Anxiety/depression	Total
ESSD	-0.168	0.015	0.059	0.203	0.475	0.209

While significant differences in item parameters have been widely identified, these may not be large enough to be of practical importance. The impact of DiF on the expected scores of the two groups across the latent health trait is shown in Appendix 17 and presented graphically in Appendix 18. The impact of DiF on expected item and utility scores accounts for the preference weighting of the EQ-5D-5L using the published EQ-5D-5L tariff (Devlin, Shah et al., 2017). Expected score standardised difference (ESSD) effect sizes have also been calculated for each item, shown in Table 15, in order to provide some guidance on where DiF may and may not have practical importance. Older adults are expected to score higher, given the same underlying HRQoL, on all items except mobility, where they are expected to score

lower, however the magnitude of the impact varies substantially. As shown in Table 15, the impact of DiF on self-care, usual activities and mobility is trivial. The impact on pain/discomfort and anxiety/depression is small but the impact on anxiety/depression approaches moderate, according to Cohen's d classification (Cohen, 1988). The effect size of the impact of DiF on measure as a whole was small. The maximum difference in EQ-5D-5L utility (Figure 15) is 0.127, 9.86% of the score range (possible score range 1 to -0.285) at 1.75 SDs below mean HRQoL, with older adults expected to score higher.

Figure 15 – Impact of DiF on expected EQ-5D-5L utility



As the EQ-5D is currently the measure required by NICE for use in the economic evaluation of healthcare interventions, it is important to explore the impact that this DiF would have on the results of economic evaluation in terms of bias in estimates of effectiveness and incremental effectiveness. Using the expected utilities of under and over 65s at each underlying level of health, we can imagine the results of an economic evaluation of a new treatment compared to a control group receiving the current standard care. Six hypothetical economic evaluations are outlined below in Table 16. In each of the six separate trials, both under and over 65s start at the same baseline level of health, according to the underlying trait measured by the IRT model. After some amount of time the control group gain some health and move 0.25 SDs up the latent trait. The treatment group gain more health and move 0.5 SDs up the latent trait. Within each treatment group, the same underlying amount of health has been gained by both age groups. However, DiF means that at the same underlying level of health, the age groups have different expected utilities. This framework allows us to

examine whether there is bias in the effectiveness and incremental effectiveness estimated in each age group.

Interestingly the bias observed is not constant across trials in either magnitude or direction of bias. In hypothetical trial 1, we see that for the same underlying gain in health, older adults receive a higher estimate of effectiveness in both the treatment and control group. These higher effectiveness estimates in both groups do not cancel each other out, as we may expect, and older adults also receive a higher level of incremental effectiveness. Starting at a higher level of baseline utility, but maintaining the same patterns in underlying health gain, hypothetical trials 2, 3 and 4 result in higher estimates of effectiveness and incremental effectiveness for younger adults. However, the difference in effectiveness and incremental effectiveness estimates between the age groups declines once we reach those individuals with above average underlying health as both age groups begin to reach the ceiling of the measure in hypothetical trials 5 and 6.

These differences in the direction of bias and magnitude of bias can be explained by the ICC curves of the EQ-5D-5L items (shown in Appendix 16), particularly those of pain/discomfort and anxiety/depression. From these we see that older adults at the very bottom of the underlying health scale, between -3 and -2.5 SDs, are much more likely to shift away from the most severe categories of pain/discomfort and anxiety/depression than younger adults. Therefore, it is likely that for a treatment which improves QoL, which treats a condition with the very high burden of illness with baseline utilities worse than dead, a higher proportion of older adults will endorse a higher response category at follow up than will younger adults. However, because older adults move away from the extreme and severe categories much quicker than younger adults, they exhaust these utility gains at very low levels of QoL and do not have as many improvements to endorse above this level. Therefore, it is likely that for a treatment which improves the QoL of patients with less burden of illness and baseline utilities better than dead, older adults will receive lower estimates of effectiveness than younger adults who have more categories with which they can signal improvements than older adults. Differences in effectiveness between age groups diminish to nothing as we reach conditions with very low burden of illness as both age groups are likely to be near the ceiling of the measure and cannot signal further improvements.

Table 16 - Expected EQ-5D-5L utilities and hypothetical effectiveness and incremental effectiveness of a new intervention compared to standard care by age group

Trial	HRQoL	Expected Utility			Treatment group				
		Under 65	Over 65	Difference between age groups					
1	-3	-0.20	-0.12	0.09	Baseline 1		younger	older	younger older
	-2.75	-0.16	-0.06	0.10	Control 1	Effectiveness	0.101	0.121	0.043 0.053
	-2.5	-0.10	0.01	0.11	Treatment 1	Incremental effect	0.058	0.067	
	-2.25	-0.03	0.09	0.12					
2	-2	0.05	0.17	0.12	Baseline 2		younger	older	younger older
	-1.75	0.15	0.27	0.13	Control 2	Effectiveness	0.216	0.203	0.098 0.100
	-1.5	0.26	0.38	0.11	Treatment 2	Incremental effect	0.118	0.104	
	-1.25	0.40	0.50	0.10					
3	-1	0.52	0.60	0.07	Baseline 3		younger	older	younger older
	-0.75	0.63	0.69	0.05	Control 3	Effectiveness	0.20	0.15	0.11 0.09
	-0.5	0.72	0.75	0.03	Treatment 3	Incremental effect	0.09	0.06	
	-0.25	0.79	0.80	0.01					
4	0	0.85	0.86	0.01	Baseline 4		younger	older	younger older
	0.25	0.89	0.91	0.01	Control 4	Effectiveness	0.078	0.081	0.045 0.049
	0.5	0.93	0.94	0.01	Treatment 4	Incremental effect	0.034	0.033	
	0.75	0.95	0.96	0.01					
5	1	0.97	0.98	0.01	Baseline 5		younger	older	younger older
	1.25	0.98	0.98	0.00	Control 5	Effectiveness	0.020	0.015	0.012 0.009
	1.5	0.99	0.99	0.00	Treatment 5	Incremental effect	0.008	0.006	
	1.75	0.99	1.00	0.00					
6	2	1.00	1.00	0.00	Baseline 6		younger	older	younger older
	2.25	1.00	1.00	0.00	Control 6	Effectiveness	0.003	0.002	0.001 0.001
	2.5	1.00	1.00	0.00	Treatment 6	Incremental effect	0.001	0.001	

SF-12v2 TLA

Fit statistics were mixed, although an improvement on the single-group model. The RMSEA mean (90% CI) was 0.114 (0.110, 0.118), above the cut-off for acceptable fit but the CFI was good at 0.987. The density plots of physical and mental health in Figure 16 show the distributions of the estimated levels of physical and mental health of under and over 65s. The distributions, particularly for physical health appear fairly normal. Physical and mental health are distributed from about 2.75 SDs below to 1.75 SDs above the mean level in under 65s. In over 65s physical health is distributed from about 3.5 SDs below to 1.5 SDs above the mean level of physical health and mental health is distributed between approximately 2.75 SDs below to 1.5 SDs above the mean level of mental health. There are no obvious substantial floors or ceilings for the measure as a whole which would limit its ability to discriminate the health of large groups of respondents.

Unstandardised discrimination parameters (Table 17) on the physical health factor range from 1.24 on energy in both groups to 2.52 for physical role in under 65s. In over 65s unstandardized discriminations in the physical health factor range from 1.38 for pain to 2.52 for physical role. On the mental health factor unstandardized discriminations range from 1.29 for mental health to 2.43 on emotional role in both groups. All items are relevant to health, with the lowest standardised discrimination being 0.78 for energy in under 65s. Differences in discrimination parameters indicate non-uniform DIF, with general health and energy being more closely related to the physical health of over 65s and pain and social activities being more closely related to the physical health of under 65s.

The difficulty parameters for some items show that they may have limited ability to discriminate the health of a wide range of respondents (Table 18). The difficulty parameters for physical functioning range from -1.01 to 0.36 SDs in under 65s and -1.00 to 0.55 SDs in over 65s. Anyone outside this range is most likely to answer at the floor or ceiling. Anyone with slightly above average health is most likely to respond in the highest category for emotional role ($b_4=0.07$), pain ($b_4=0.43$ in under 65s and $b_4=0.24$ in over 65s) and social activities ($b_4=0.23$ in under 65s and $b_4=-0.17$ in over 65s) meaning they provide little information for respondents with above average health. Difficulty parameters for physical functioning were lower in under 65s, indicating that older adults are more likely than younger adults to respond lower to this scale and more likely to signal problems with physical functioning than a younger adult with the same underlying level of physical health. However, difficulty parameters

for general health, pain, mental health, energy, social activities and difficulty parameters 1-4 of physical role were all higher in under 65s, suggesting older adults would be more likely to respond higher to these items than younger adults with the same level of underlying health.

Table 17 – SF-12v2 TLA Factor structures, unstandardised discrimination parameters and absolute model fit statistics

	Unstandardised discrimination parameters (SEs)			
	Under 65s		Over 65s	
SF-12v2 TLA	Physical	Emotional	Physical	Emotional
1.General health	1.46 (0.03)		1.62 (0.06)	
2/3. Physical functioning	2.16 (0.05)		2.16 (0.05)	
4/5.Physical role	2.52 (0.06)		2.52 (0.06)	
6/7.Emotional role		2.43 (0.09)		2.43 (0.09)
8.Pain	1.66 (0.04)		1.38 (0.05)	
9/11.Mental Health		1.29 (0.03)		1.29 (0.03)
10.Energy	1.24 (0.03)		1.44 (0.05)	
12.Social activities	2.00 (0.05)		1.80 (0.07)	
Factor mean	0	0	-0.424 (0.03)	-0.13 (0.03)
Factor variance	1	1	0.978 (0.06)	0.727 (0.04)
Factor correlation	0.819 (0.01)		0.825 (0.06)	
Model Fit	RMSEA mean (90% CI)		CFI	
	0.114 (0.110, 0.118)		0.987	

Discrimination parameters in bold exhibit non-uniform DiF as they were found to differ between under and over 65s.

Constraining the unstandardized residual variances of both groups to 1 did not significantly impact the fit of the model (*p*-value=0.25) indicating that the amount of item variance not accounted for by the factor was the same across age groups. The factor means differed between the two age groups, as shown in Table 17. The mean level of physical health is 0.42 SDs lower in the over 65 group (*p*-value<0.001) while the mean level of mental health is 0.13 SDs lower in over 65s (*p*-value<0.001). The factor variances were found to differ between groups. As shown in Table 17, the over 65 group were slightly less variable in physical health with a factor variance of 0.98, compared to the constrained factor variance of 1 in under 65s, while over 65s were less variable in mental health than under 65s, with a mental health factor variance of 0.73. Factor invariance was then tested to examine whether constraining the factor variances to equal 1 in both groups significantly impacted the model. The DIFFTEST was significant (*p*-value<0.001), suggesting that constraining the factor variance to 1 in both groups did significantly impact model fit and group factor variances should be allowed to differ.

Figure 16 – Density plot of predicted Physical Health from the SF-12v2 TLA

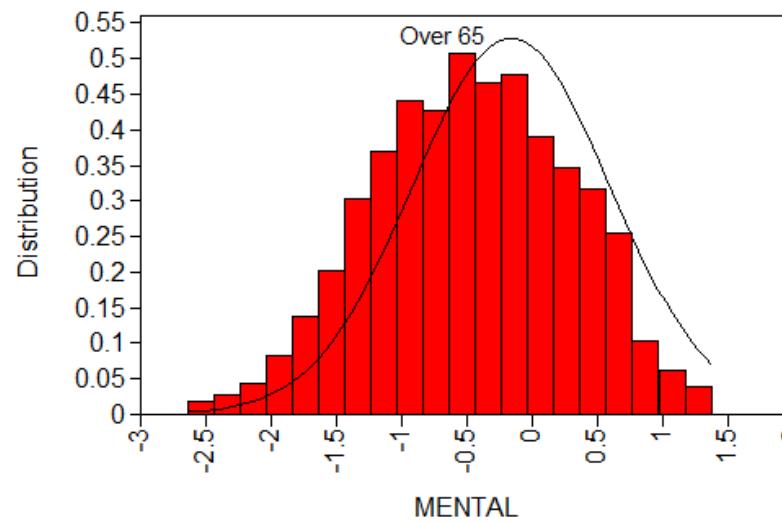
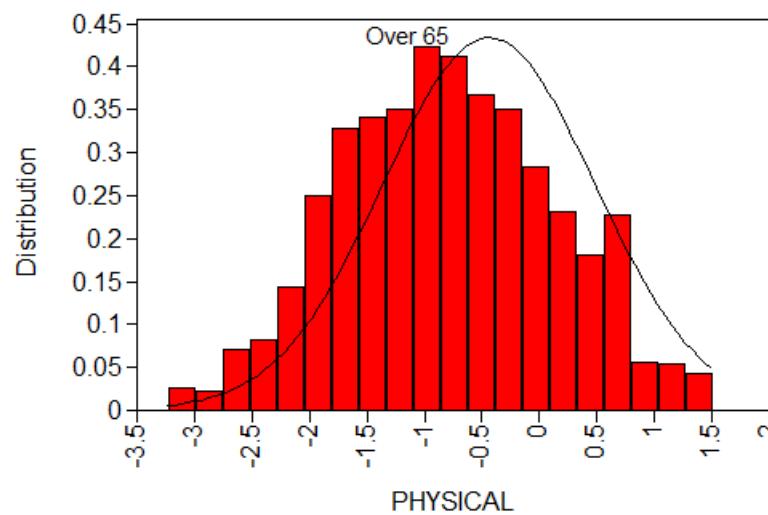
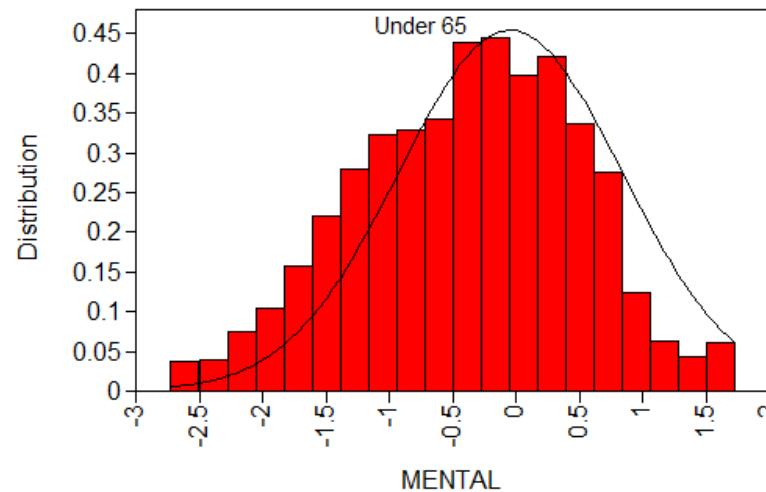
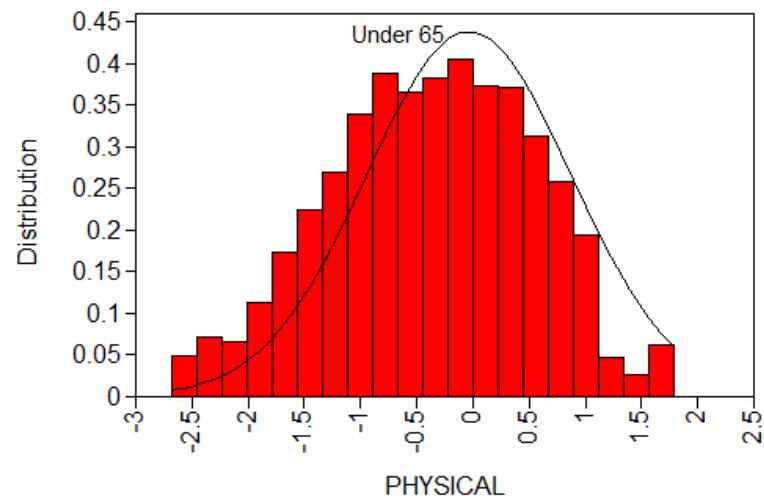


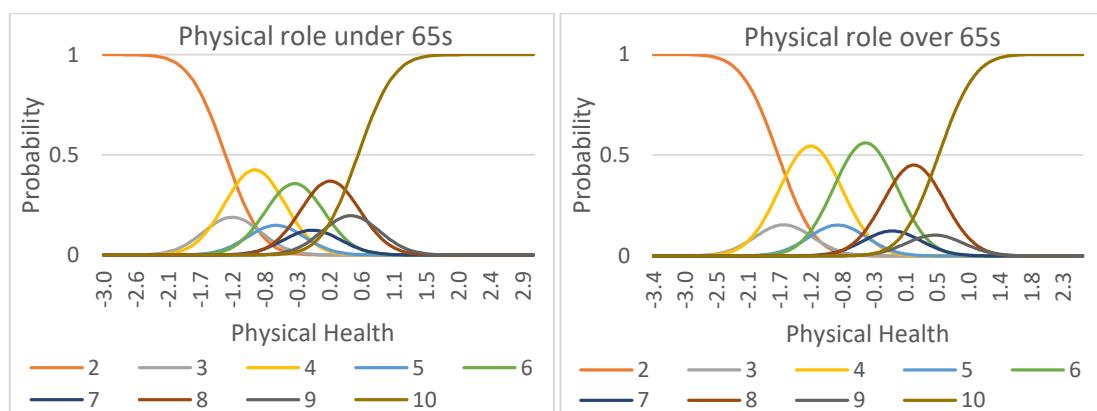
Table 18 – SF-12v2 TLA difficulty parameters

Difficulty Parameters (SEs)																
Under 65s								Over 65s								
b1	b2	b3	b4	b5	b6	b7	b8	SF-12v2 TLA	b1	b2	b3	b4	b5	b6	b7	b8
-1.43 (0.05)	-0.52 (0.04)	0.38 (0.04)	1.59 (0.05)	N/A	N/A	N/A	N/A	1.General health	-1.84 (0.07)	-0.66 (0.05)	0.34 (0.04)	1.62 (0.09)	N/A	N/A	N/A	N/A
-1.01 (0.06)	-0.63 (0.06)	-0.07 (0.05)	0.36 (0.05)	N/A	N/A	N/A	N/A	2/3. Physical functioning	-1.00 (0.07)	-0.55 (0.06)	0.12 (0.06)	0.56 (0.07)	N/A	N/A	N/A	N/A
-1.30 (0.08)	-1.11 (0.07)	-0.66 (0.06)	-0.52 (0.06)	-0.15 (0.05)	-0.03 (0.05)	-0.36 (0.06)	0.55 (0.06)	4/5.Physical role	-1.66 (0.10)	-1.51 (0.10)	-0.92 (0.08)	-0.76 (0.08)	-0.15 (0.05)	-0.03 (0.05)	0.45 (0.07)	0.55 (0.06)
-1.73 (0.12)	-1.56 (0.11)	-1.13 (0.09)	-0.97 (0.09)	-0.57 (0.07)	-0.43 (0.06)	-0.09 (0.06)	0.07 (0.05)	6/7.Emotional role	-1.73 (0.12)	-1.56 (0.11)	-1.13 (0.09)	-0.97 (0.09)	-0.57 (0.07)	-0.43 (0.06)	-0.09 (0.06)	0.07 (0.05)
-1.59 (0.06)	-0.81 (0.05)	-0.32 (0.04)	0.43 (0.04)	N/A	N/A	N/A	N/A	8.Pain	-2.31 (0.07)	-1.16 (0.05)	-0.57 (0.04)	0.24 (0.04)	N/A	N/A	N/A	N/A
-2.50 (0.08)	-1.97 (0.06)	-1.45 (0.05)	-1.01 (0.04)	-0.33 (0.04)	0.17 (0.03)	-0.81 (0.04)	1.88 (0.06)	9/11.Mental Health	-2.90 (0.12)	-2.45 (0.09)	-1.96 (0.06)	-1.50 (0.05)	-0.81 (0.05)	-0.31 (0.04)	0.30 (0.04)	1.17 (0.06)
-1.28 (0.04)	-0.45 (0.03)	0.43 (0.04)	1.98 (0.04)	N/A	N/A	N/A	N/A	10.Energy	-1.52 (0.06)	-0.70 (0.05)	0.21 (0.04)	1.70 (0.05)	N/A	N/A	N/A	N/A
-1.44 (0.06)	-0.84 (0.05)	-0.24 (0.05)	0.23 (0.05)	N/A	N/A	N/A	N/A	12.Social activities	-1.91 (0.07)	-1.25 (0.07)	-0.61 (0.06)	-0.17 (0.05)	N/A	N/A	N/A	N/A

Where b1 represents the amount of QoL required to have a 50% probability of responding in the category signalling the lowest level of QoL and 50% chance of responding higher. Difficulty parameters in bold exhibit DiF as they were found to differ between under and over 65s

As shown in the ICCs for each item in Appendix 16, levels mostly behave as expected for single items. For pain, there is only a very small range over which “moderately” is the most likely choice and for social activities “a little of the time” is never the most probable choice in over 65s and only most likely over a very small range of under 65s, indicating that these categories may be underused or somewhat indistinct from neighbouring categories. The ICCs for the four super items displayed an unanticipated pattern. Consistently across the four super items, even numbered scores dominated, with odd scores very rarely having a range over which they were the most likely option to be chosen. An example of this pattern is displayed in Figure 17. While unanticipated and different to the usual ICC pattern, this can likely be explained by the fact that these item pairs are displayed together in clusters, with identical response options within each cluster. Therefore, there may be a tendency for respondents to answer the same response to each item within the pair. Answering the same response will give even scores. This is particularly likely within the role pairs as the response options are identical, the items have very similar content and they likely require a similar level of functioning to one another. In fact, we see from the ICCs that the dominance of even numbered responses is far stronger in the two pairs of role items.

Figure 17 – ICCs for SF12v2 physical role super item



Physical role provides the most information about the physical health of both groups, followed by physical functioning, while energy, pain and general health provide the lower levels of information in this factor. Emotional role dominates the information

provided about the mental health of respondents in the mental factor (Figure 18). Total information (Figure 19) for the physical health factor is higher in under 65s, but internal consistency is good (Cronbach's alpha \geq 0.8) for a broader range of over 65s (-2.6 to +1.8 SDs vs -2.2 to 1.4 SDs about the mean). Total information for the mental health factor shows good internal reliability across a range of approximately -2.2 to 0.5 SDs about the mean. This suggests that the mental health factor will struggle to precisely discriminate those with above average health.

Figure 18 – SF-12v2 TLA item information by factor and age group

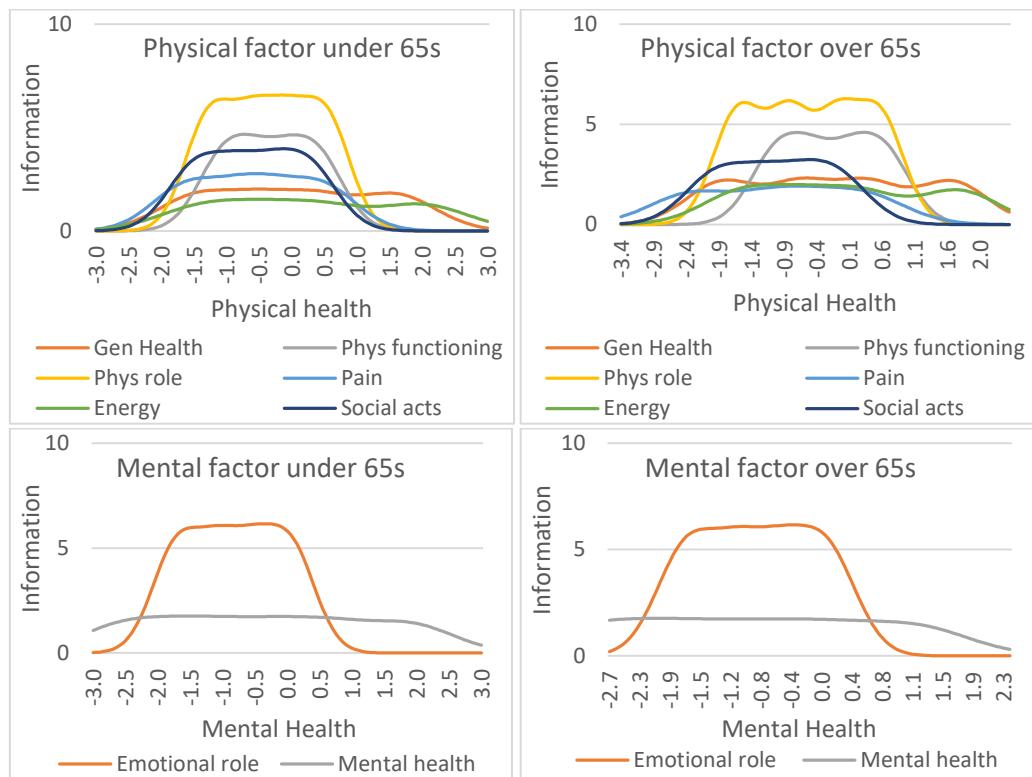
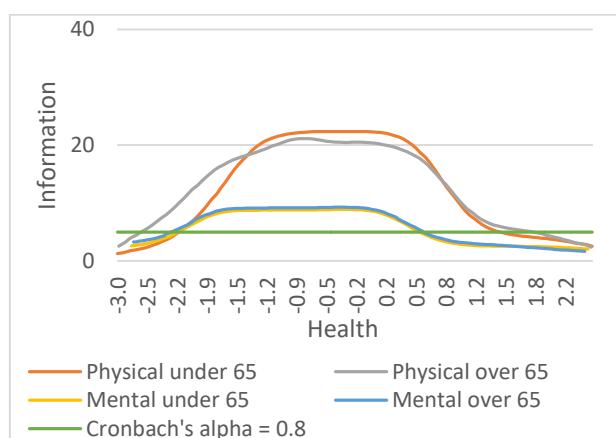


Figure 19 – SF-12v2 TLA Total Information

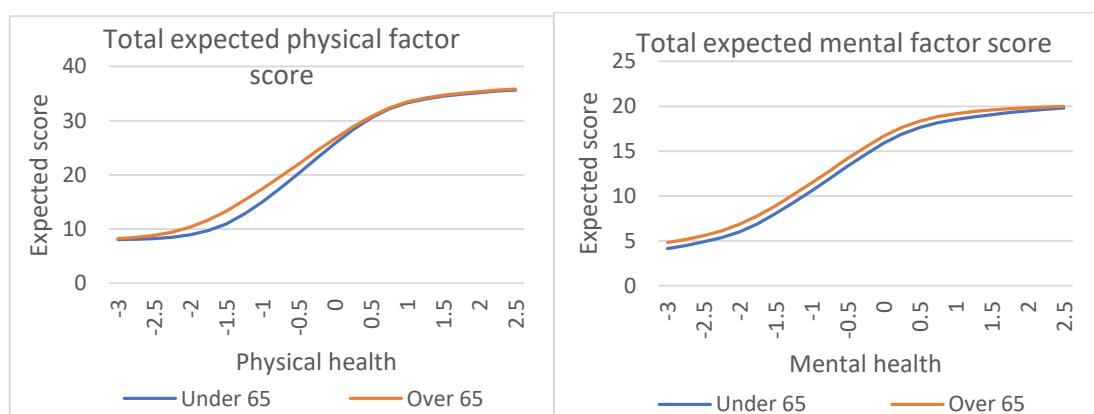


The effect sizes for item level DiF are shown in Table 19. DiF had a moderate impact on the expected scores of the mental health super item, with older adults expected to score higher than younger adults with the same underlying health. DiF had a small impact on physical role, pain, energy and social activities, again with older adults expected to score higher and a trivial impact on the remaining items. The effect size for the impact of DiF on the physical scale as a whole was moderate, while for the mental scale it was small. Item level expected scores across the underlying health scale in older and younger adults can be seen in tables in Appendix 17 and graphically in Appendix 18. The difference in expected factor scores is shown in Figure 20. The maximum difference in the total expected physical health factor score was 9.10% of the possible score range (possible score range = 28) at 1.25 SDs below mean health, where older respondents were expected to score higher, given the same underlying level of health as under 65s (Figure 36). The difference in the total expected mental health factor score was 5.65% of the possible score range (possible score range = 16) at 1.25 SDs below mean health, where again, older respondents were expected to score higher.

Table 19 – SF-12v2 TLA Item level and scale level DiF effect size

Item	General health	Physical Functioning	Physical role	Emotional role	Pain
ESSD	0.154	-0.132	0.233	0.000	0.338
Item	Mental Health	Energy	Social activities	Total physical	Total Mental
ESSD	0.715	0.231	0.444	0.505	0.436

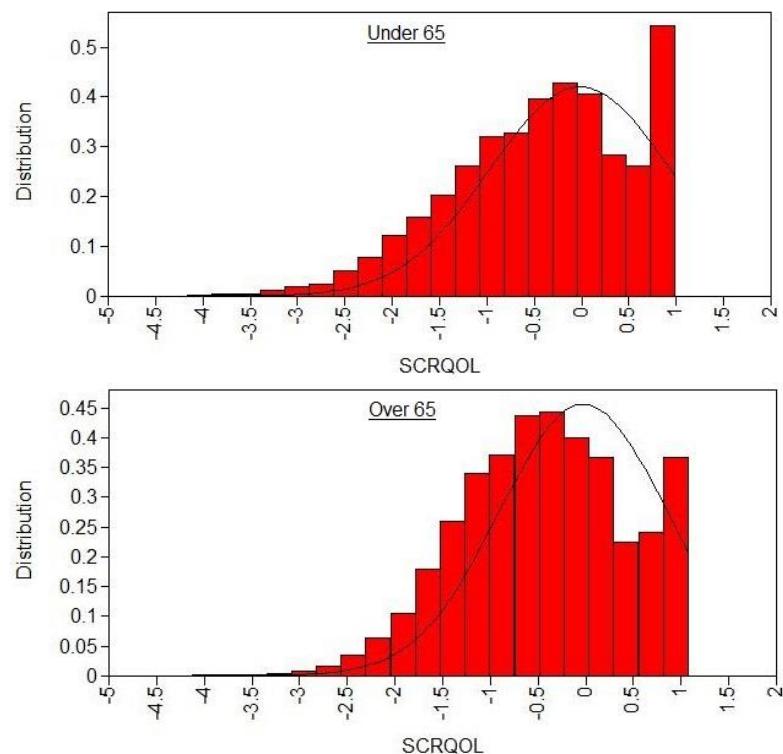
Figure 20 – Impact of DiF on SF-12v2 TLA factor score



ASCOT

The absolute fit statistics were good with an RMSEA of 0.059 (0.058, 0.061) and CFI=0.981. The density plots of SCRQoL in Figure 21 show the distributions of the estimated levels SCRQoL in under and over 65s. Both age groups are distributed between approximately 3.5 SDs below and 1 SD above the mean level of SCRQoL and substantially negatively skewed. There is a substantial ceiling in ASCOTs ability to discriminate the SCRQoL of those respondents above 1 SD above the mean level of SCRQoL, shown by the large peak at the top end of the distribution.

Figure 21 – Density plot of predicted SCRQoL from the ASCOT



Unstandardised discriminations (Table 20) range from 0.63 for item 8 dignity to 1.32 for item occupation in both groups. All items are relevant to SCRQoL, with the lowest standardised discrimination being 0.534 for dignity in over 65s. No item redundancy was suggested by the discrimination parameters. Non-uniform DiF in the discrimination parameters suggested that control and personal cleanliness/comfort were more closely related to the SCRQoL of older respondents and accommodation cleanliness/comfort and safety were less closely related to the SCRQoL of older adults than younger adults.

Table 20 – ASCOT factor structures, unstandardised discrimination parameters and absolute model fit statistics

	Unstandardised discrimination Parameters (SEs)	
	Under 65s	Over 65s
ASCOT	SCRQoL	SCRQoL
1.Control	0.83 (0.02)	1.07 (0.01)
2.Personal clean	1.04 (0.02)	1.12 (0.02)
3.Food/Drink	0.87 (0.01)	0.87 (0.01)
4.Safety	0.98 (0.02)	0.78 (0.01)
5.Social participation	1.06 (0.01)	1.06 (0.01)
6.Occupation	1.32 (0.02)	1.32 (0.02)
7.Accommodation clean	0.99 (0.02)	0.89 (0.02)
8.Dignity	0.63 (0.01)	0.63 (0.01)
Factor mean	0.039 (0.02)	0
Factor variance	1.24 (0.03)	1
Model Fit	RMSEA mean (90% CI)	CFI
	0.059 (0.058, 0.061)	0.981

Discrimination parameters in bold exhibit non-uniform DiF as they were found to differ between under and over 65s. Clean=cleanliness and comfort

Low b3 parameters (Table 21) across all items mean individuals above 0.635 SDs above mean SCRQoL will most likely report no problems on all questions, meaning the measure will have very low power to discriminate above this level. Items dignity, accommodation, food/drink and personal cleanliness/comfort all had b1s more than 3 SDs below mean SCRQoL. We would expect very few people to have a SCRQoL below this level. Therefore, it is unlikely that respondents will endorse category 1, which signals the lowest level of SCRQoL, on these items.

Difficulty parameters (Table 21) for control and occupation were always higher for over 65s, signalling uniform DiF, resulting in older adults being more likely to respond lower to these items than younger adults, whereas younger adults were more likely to respond lower for safety. For food/drink, social, accommodation and dignity; b1s were all lower for over 65s but b3s were all higher for over 65s. This suggests that older adults with low levels of SCRQoL are more likely to respond higher than younger adults, while older adults with higher levels of SCRQoL choosing between b2 and b3 are more likely to respond lower. This will compress the scores of older adults compared to younger adults with the same SCRQoL.

Table 21 – ASCOT difficulty parameters

Difficulty Parameters (SEs)							
Under 65				Over 65			
b1	b2	b3	ASCOT	b1	b2	b3	
-2.97 (0.03)	-1.42 (0.02.)	0.64 (0.02)	1.Control	-2.08 (0.02)	-0.88 (0.02)	0.72 (0.01)	
-3.25 (0.05)	-2.17 (0.03)	-0.46 (0.02)	2.Personal clean	-3.40 (0.05)	-2.22 (0.03)	-0.08 (0.01)	
-3.40 (0.04)	-2.44 (0.02)	-0.61 (0.02)	3.Food/Drink	-3.62 (0.04)	-2.44 (0.02)	-0.48 (0.01)	
-2.90 (0.03)	-2.15 (0.02)	-0.61 (0.01)	4.Safety	-3.66 (0.03)	-2.71 (0.02)	-0.78 (0.01)	
-2.18 (0.03)	-1.07 (0.01)	0.08 (0.02)	5.Social participation	-2.36 (0.02)	-1.07 (0.01)	0.32 (0.01)	
-2.31 (0.04)	-0.86 (0.02)	0.19 (0.02)	6.Occupation	-1.67 (0.02)	-0.43 (0.02)	0.70 (0.02)	
-3.38 (0.05)	-2.29 (0.03)	-0.64 (0.02)	7.Accommodation clean	-4.07 (0.06)	-2.70 (0.03)	-0.56 (0.01)	
-4.12 (0.03)	-2.53 (0.01)	-0.86 (0.02)	8.Dignity	-4.36 (0.03)	-2.53 (0.01)	-0.37 (0.01)	

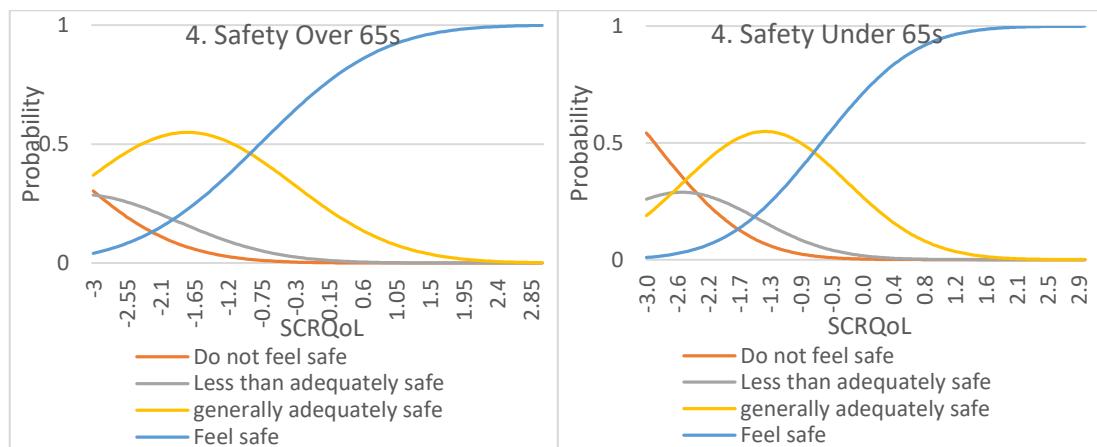
Where b1 represents the amount of QoL required to have a 50% probability of responding in the category signalling the lowest level of QoL and 50% chance of responding higher. Difficulty parameters in bold exhibit uniform DiF as they were found to differ between under and over 65s. Clean=cleanliness and comfort

Constraining the unstandardized residual variances of both groups to 1 did not significantly impact the fit of the model ($p\text{-value}=0.259$) indicating that the amount of item variance not accounted for by the factor was the same across groups. The factor mean differed between the two age groups, as shown in Table 20. The mean level of SCRQoL is 0.039 SDs higher in the under 65 group. This difference was significant ($P\text{-value}=0.024$). The factor variance was found to differ between groups. As shown in Table 20, the under 65 group were more variable in SCRQoL with a factor variance of 1.24, compared to the constrained factor variance of 1 in over 65s. Factor invariance was then tested to examine whether constraining the factor variance to equal 1 in both groups significantly impacted the model. The DIFFTEST was significant ($p\text{-value}<0.000$), suggesting that constraining the factor variance to 1 in both groups did significantly impact model fit and group factor variances do differ significantly.

The ICCs for each item in each age group are shown in Appendix 16. There is an issue with response option 2 for safety (Figure 22), which is never the most likely response option in either group. There is a potential problem with the wording of this

level. This response option reads “I feel less than adequately safe” while its neighbouring categories 1 and 3 read “I don’t feel at all safe” and “generally I feel adequately safe, but not as safe as I would like”. Respondents potentially struggle to distinguish the middle category clearly from its neighbours. This category may need rewording.

Figure 22 – Examples of problematic item characteristic curves from the ASCOT



Occupation provides the highest level of information and dignity the least (Figure 23). All item information curves sharply decline by 1 SD above the mean. The measure has good internal reliability (Figure 24) from -3 SDs in both groups to approximately 1.3 SDs above the mean in over 65s and 0.8 SDs in under 65s.

Figure 23 – ASCOT item information by age group

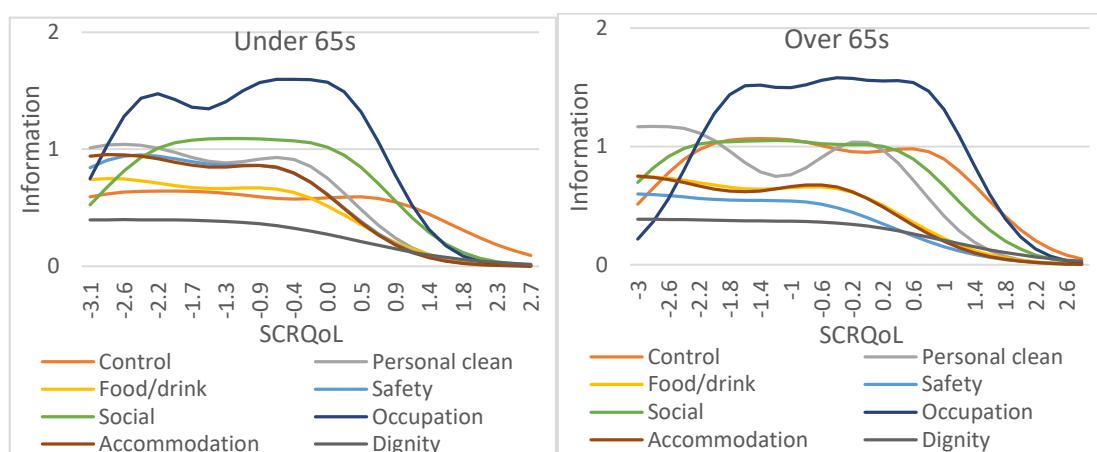
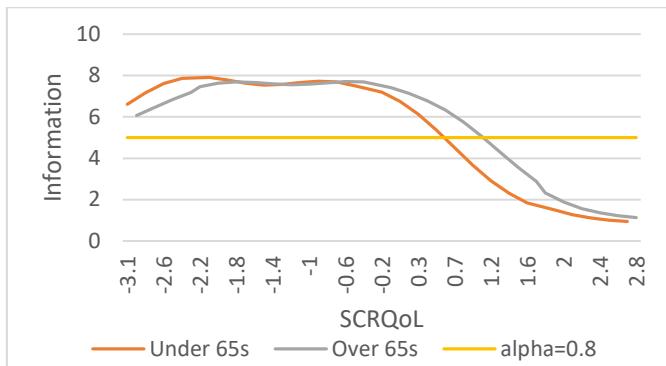


Figure 24 – ASCOT total information



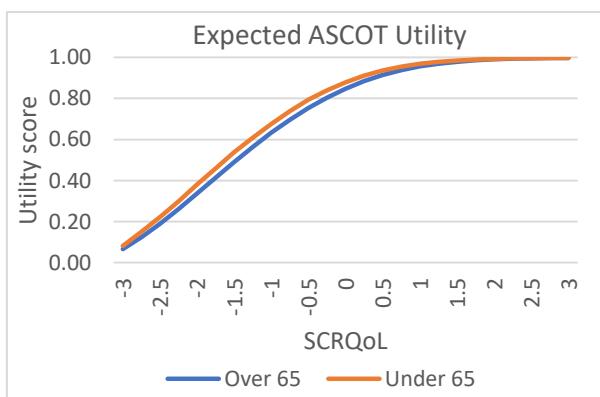
As shown in Table 22, the effect size for the impact of DiF on item expected scores was trivial for the majority of items, but was small for control and occupation, with older adults expected to score lower to both these items than a younger adult with the same underlying SCRQoL. The effect size for the measure as a whole was also small. The impact of DiF on ASCOT items and utility, accounting for the preference weighting of the ASCOT are displayed in Appendices 17 and 18. DiF impact on the ASCOT utility (Figure 25) (possible score range 1 to -0.17) reached a maximum of 4.1% of the score range in individuals 1.5 SDs below the mean SCRQoL level, where older adults were expected to score slightly lower.

Table 22 – ASCOT Item level DiF effect size

Item	Control	Personal clean	Food/ Drink	Safety	Social	Occupation	Accommodation	Dignity	Total
ESSD	0.206	0.14	0.043	-0.068	0.07	0.394	0.009	0.166	0.37

Clean=cleanliness and comfort

Figure 25 – Impact of DiF on ASCOT utility score



WEMWBS

Fit statistics were mixed, although an improvement on the single factor model. Mean RMSEA (90% CI) was 0.090 (0.087, 0.093), above the cut-off for acceptable fit but CFI was good at 0.981. The density plots of internal and external wellbeing in Figure 26 shows the distributions of the estimated levels of internal and external wellbeing in under and over 65s. They are all slightly negatively skewed. Internal wellbeing is distributed between approximately 4 SDs below and 2 SDs above the mean level in under and over 65s while external wellbeing is distributed between the same levels in under 65s and between approximately 3 SDs below and 1.75 SDs above the mean level in over 65s. There are no obvious substantial measure ceiling and floor effects limiting the ability to discriminate the full range of wellbeing of respondents.

Discrimination parameters (Table 23) on the external factor range from 1.36 for feeling close to other people to 0.85 for feeling optimistic about the future in under 65s and 1.47 for feeling useful to 0.85 for feeling optimistic about the future in over 65s. Discriminations on the internal factor ranged from 0.86 for having energy to spare to 2.08 for feeling good about oneself in both groups. All items are relevant to mental wellbeing, with the lowest standardised discrimination being 0.59 for feeling optimistic about the future in over 65s. Two similar items; feeling good about oneself and feeling confident have particularly high discriminations suggesting item redundancy. The discrimination parameters signalled non-uniform DiF for feeling useful, interested in other people and able to make up your own mind about things, as they were all higher in over 65s suggesting they are more closely related to the mental wellbeing of over 65s than under 65s.

Figure 26 – Density plot of predicted internal wellbeing from the WEMWBS

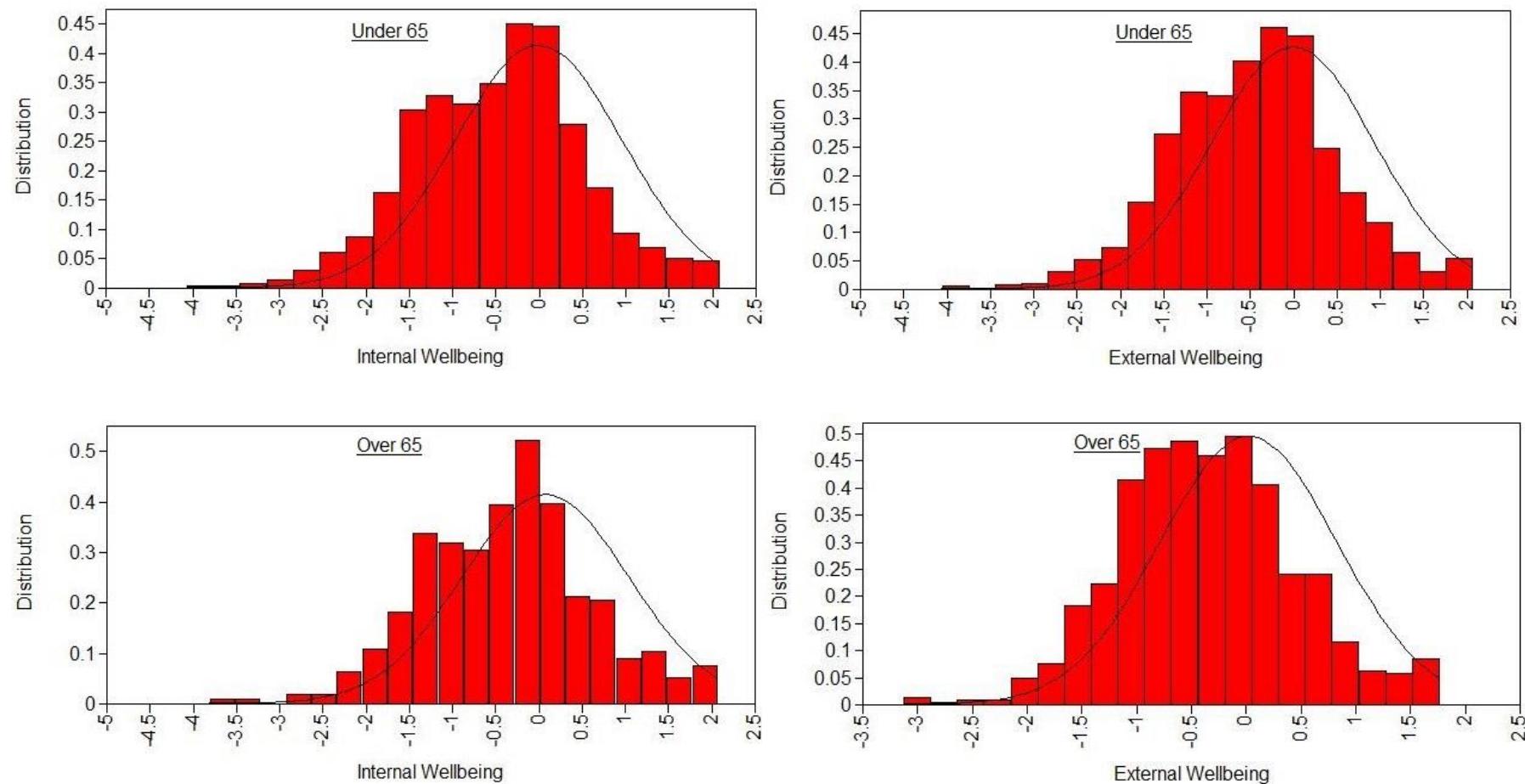


Table 23 – WEMWBS Factor structures, unstandardised discrimination parameters and absolute model fit statistics

WEMWBS	Unstandardised discrimination parameters (SEs)			
	Under 65s		Over 65s	
	External	Internal	External	Internal
1.Optimistic about future	0.85 (0.02)		0.85 (0.02)	
2.Useful	1.13 (0.03)		1.47 (0.08)	
3.Relaxed		1.21 (0.03)		1.21 (0.03)
4.Interested other people	0.88 (0.03)		1.03 (0.06)	
5.Energy to spare		0.86 (0.02)		0.86 (0.02)
6.Deal problems		1.45 (0.03)		1.45 (0.03)
7.Thinking clearly		1.46 (0.03)		1.46 (0.03)
8.Feel good about self		2.08 (0.04)		2.08 (0.04)
9.Close other people	1.36 (0.04)		1.36 (0.04)	
10.Confident		2.07 (0.05)		2.07 (0.05)
11.Make own mind		1.18 (0.03)		1.35 (0.06)
12.Loved	1.12 (0.03)		1.12 (0.03)	
13.Interested new things	1.18 (0.03)		1.18 (0.03)	
14.Cheerful		1.81 (0.04)		1.81 (0.04)
Factor mean	0	0	0.023 (0.04)	0.079 (0.04)
Factor variance	1	1	0.742 (0.06)	1.002 (0.06)
Factor correlation	0.895 (0.01)		0.731 (0.05)	
Model Fit	RMSEA mean (90% CI)		CFI	
	0.09 (0.087, 0.093)		0.973	

Discrimination parameters in bold exhibit non-uniform DiF as they were found to differ between under and over 65s

Difficulty parameters for each item (Table 24) tend to cover a broad range of underlying wellbeing, suggesting the items can discriminate the wellbeing across respondents. All difficulty parameters for feeling optimistic and having energy to spare exhibited uniform DiF, as they were higher in older adults, suggesting they were more likely to respond lower than a younger adult at any given level of wellbeing. Among the remaining difficulty parameters which differ between age groups, older adults tended to be more likely to respond lower to feeling useful and interested in new things but higher to feeling relaxed and confident.

Table 24 – WEMWBS difficulty parameters

Difficulty Parameters (SEs)									
Under 65s					Over 65s				
b1	b2	b3	b4	WEMWBS	b1	b2	b3	b4	
-2.64 (0.05)	-1.57 (0.04)	0.09 (0.03)	1.87 (0.04)	1.Optimistic about future	-2.11 (0.08)	-1.09 (0.06)	0.52 (0.05)	1.91 (0.07)	
-2.50 (0.06)	-1.74 (0.05)	-0.29 (0.04)	1.42 (0.04)	2.Useful	-1.92 (0.06)	-1.33 (0.05)	-0.02 (0.07)	1.09 (0.04)	
-2.54 (0.07)	-1.27 (0.05)	0.26 (0.04)	1.86 (0.06)	3.Relaxed	-2.54 (0.07)	-1.77 (0.09)	-0.03 (0.06)	1.48 (0.08)	
-2.92 (0.06)	-1.83 (0.04)	-0.28 (0.03)	1.64 (0.04)	4.Interested other people	-2.51 (0.06)	-1.87 (0.08)	-0.24 (0.03)	1.11 (0.06)	
-2.46 (0.05)	-0.82 (0.03)	0.92 (0.04)	2.51 (0.05)	5.Energy to spare	-1.89 (0.07)	-0.62 (0.06)	1.22 (0.06)	2.83 (0.10)	
-2.60 (0.08)	-1.82 (0.06)	-0.33 (0.04)	1.33 (0.05)	6.Dealing problems	-2.60 (0.08)	-1.82 (0.06)	-0.33 (0.04)	1.06 (0.08)	
-2.76 (0.10)	-1.96 (0.06)	-0.64 (0.04)	0.98 (0.05)	7.Thinking clearly	-2.76 (0.10)	-1.96 (0.06)	-0.64 (0.04)	0.64 (0.07)	
-2.36 (0.12)	-1.43 (0.07)	-0.11 (0.06)	1.27 (0.07)	8.Feel good about self	-2.36 (0.12)	-1.43 (0.07)	0.01 (0.09)	1.12 (0.11)	
-2.61 (0.09)	-1.64 (0.05)	-0.28 (0.04)	1.21 (0.05)	9.Close to other people	-2.61 (0.09)	-1.64 (0.05)	-0.28 (0.04)	1.03 (0.07)	
-2.32 (0.13)	-1.42 (0.08)	-0.16 (0.06)	1.21 (0.07)	10.Confident	-2.61 (0.29)	-1.53 (0.13)	-0.07 (0.09)	1.10 (0.11)	
-3.33 (0.12)	-2.28 (0.07)	-1.00 (0.04)	0.69 (0.04)	11.Make up own mind	-2.92 (0.12)	-2.31 (0.13)	-0.87 (0.04)	0.42 (0.07)	
-2.84 (0.08)	-1.98 (0.05)	-0.80 (0.04)	0.40 (0.03)	12.Loved	-2.84 (0.08)	-1.98 (0.05)	-0.80 (0.04)	0.40 (0.03)	
-2.58 (0.08)	-1.59 (0.05)	-0.27 (0.04)	1.18 (0.04)	13.Interested new things	-2.30 (0.12)	-1.36 (0.07)	-0.05 (0.06)	1.18 (0.04)	
-2.62 (0.13)	-1.67 (0.07)	-0.34 (0.05)	1.26 (0.06)	14.Cheerful	-2.62 (0.13)	-1.67 (0.07)	-0.34 (0.05)	1.09 (0.10)	

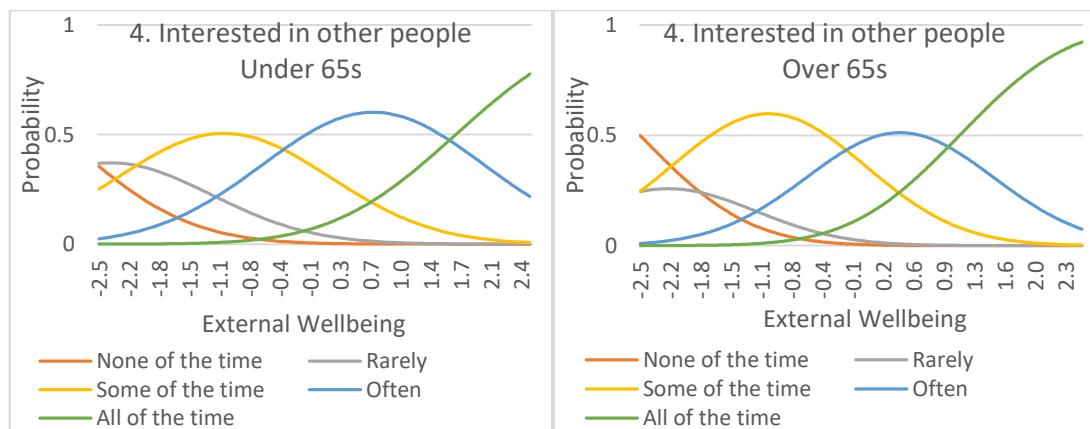
Where b1 represents the amount of QoL required to have a 50% probability of responding in the category signalling the lowest level of QoL and 50% chance of responding higher. Difficulty parameters in bold exhibit uniform DiF as they were found to differ between under and over 65s

Constraining the unstandardized residual variances of both groups to 1 did significantly impact the fit of the model ($p\text{-value}=0.036$) indicating that the amount of item variance not accounted for by the factor was not the same across age groups. The residual variance of feeling good about oneself was found to be higher in the over 65 group (1.322 vs 1). Once this was freed, constraining the remaining residual variances did not significantly impact the model ($p\text{-value}=0.124$). The factor means differed between the two age groups, as shown in Table 23. The mean level of

external wellbeing is 0.023 SDs higher in the over 65 group while the mean level of internal wellbeing is 0.079 SDs higher in over 65s. The difference in means was insignificant for both factors (external factor p-value=0.555, internal factor p-value=0.076). The factor variances were found to differ between groups. As shown in Table 23, the over 65 group were less variable in external wellbeing with a factor variance of 0.742, compared to the constrained factor variance of 1 in under 65s, while over 65s were slightly more variable in internal wellbeing than under 65s, with a factor variance of 1.002. Factor invariance was then tested to examine whether constraining the factor variances to equal 1 in both groups significantly impacted the model. The DIFFTEST was significant (p-value=0.005), suggesting that constraining the factor variance to 1 in both groups did significantly impact model fit and group factor variances should be allowed to differ.

The levels are largely used appropriately, as shown in the ICCs for each item and age group in Appendix 16. There may be an issue with the “rarely” category for some items, particularly for feeling optimistic, useful, relaxed, interested in other people and able to make up own mind; especially in older respondents where for items useful, interested in other people (Figure 27) and able to make up own mind “rarely” is never the most likely option. This suggests it is being underused, either because it is not understood, or it is indistinct from its neighbouring categories.

Figure 27 – Examples of problematic item characteristic curves from the WEMWBS



Feeling close to other people and useful provided the most information about mental wellbeing in under and over 65s respectively in the external factor (Figure 28). In the internal factor feeling good about oneself, confident and cheerful provided most information in both groups, potentially partially due to similarity between these items

suggested by inflated discriminations and MIs. The internal reliability of the external and internal factors was good up to approximately 2.1 SDs (Figure 29).

Figure 28 – WEMWBS item information by factor and age group

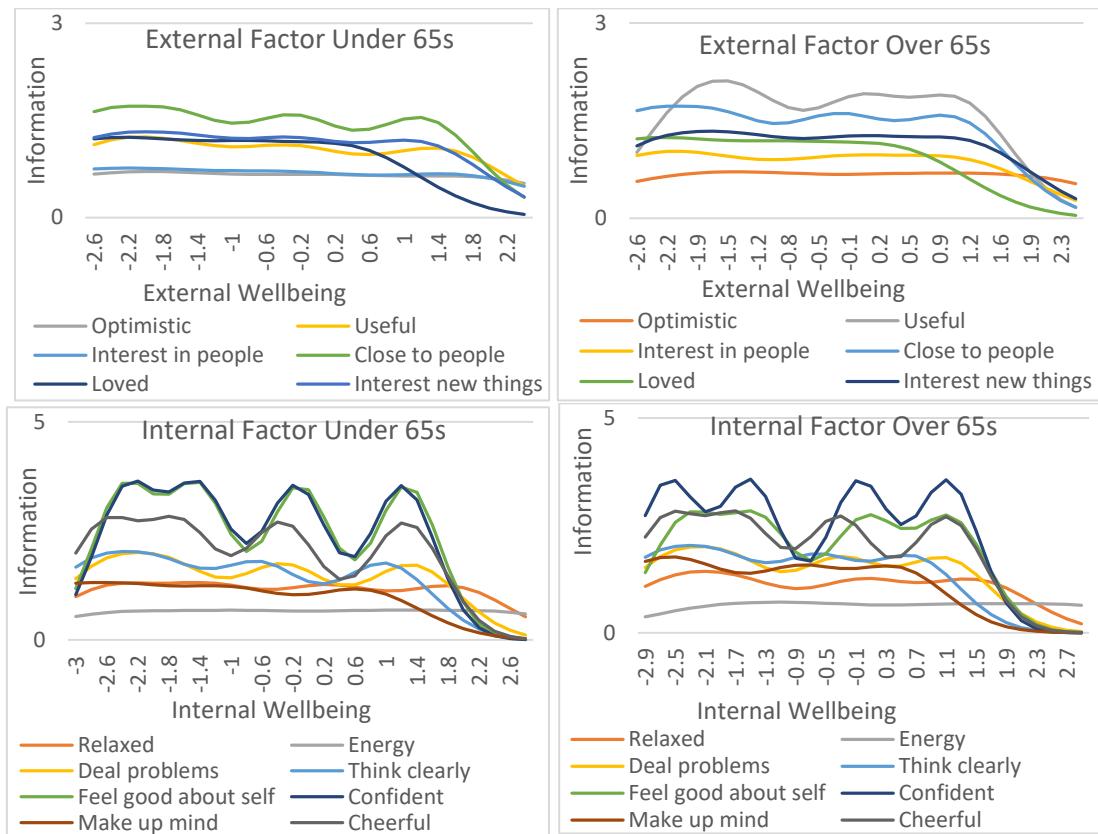


Figure 29 – WEMWBS total information

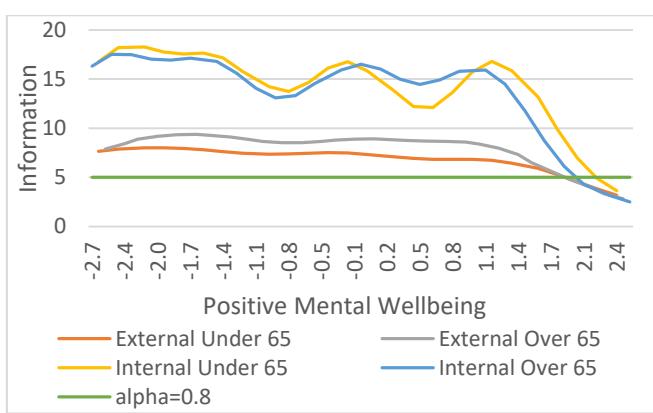
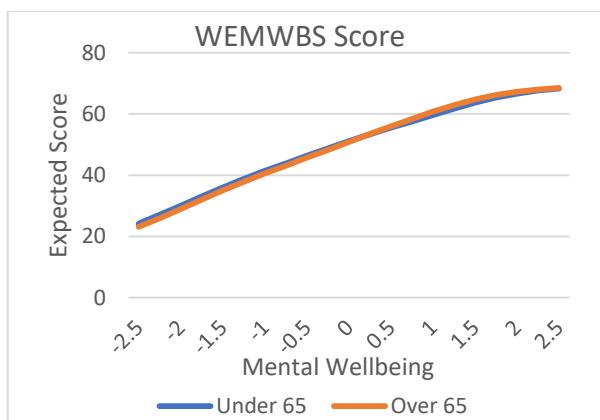


Table 25 – WEMWBS Item level DiF effect size

Item	Optimistic future	Useful	Relaxed	Interest people	Energy	Dealing problems	Thinking clearly	Feel good self
ESSD	-0.345	-0.128	0.278	0.121	-0.239	0.072	0.119	-0.024
Item	Close people	Confident	Make up mind	Loved	Interest things	Cheerful	Total Internal	Total External
ESSD	0.05	0.014	0.08	-0.002	-0.162	0.044	0.12	-0.192

As shown in Table 25, the effect size for the impact of DiF on expected items scores was mostly trivial. However, the effect size for the impact of DiF was small for feeling optimistic about the future, relaxed and having energy to spare. Amongst these three items, older adults were expected to score lower for feeling optimistic and having energy to spare than a younger adult with the same underlying wellbeing and higher for feeling relaxed. The effect size for the impact on the scale scores for the two factors was trivial for the internal factor and almost small for the internal factor, however these effects mostly cancelled each other out with older adults expected to score higher in the internal factor but lower in the external. The impact of DiF on the total WEMWBS expected scores was low, as shown in Appendices 17 and 18 and in Figure 30. Total DiF impact remained below 2% of the possible score range (56) across the entire range of wellbeing tested (Figure 33).

Figure 30 – Impact of DiF on WEMWBS score



ONS-4

Model fit statistics were good with acceptable mean RMSEA (90% CI) = 0.055 (0.044, 0.066) and good CFI=0.999. The density plots of wellbeing in Figure 31 show the distributions of the estimated levels of wellbeing in the sample of under and over 65s. Estimated wellbeing is distributed between approximately 2.5 SDs below to 1.5 SDs above mean wellbeing in under 65s and between approximately 2.5 SDs below to 1.1 SDs above mean wellbeing in over 65s. Both distributions are slightly negatively skewed. There is evidence of a ceiling in the ONS-4's ability to discriminate the wellbeing of respondents at the top end of the wellbeing scale, shown by the high proportion of respondents being estimated at the top of the distribution. The measure is unable to discriminate the wellbeing of respondents above this level.

Unstandardised discrimination parameters ranged from 1.024 and 1.188 for anxiety in under and over 65s respectively to 3.017 for happiness in both groups (Table 26). This suggests that for both groups, happiness is the closest related to wellbeing and anxiety the least. All items were found to be related to wellbeing with the lowest standardised discrimination being 0.697 on anxiety in over 65s. The only item to exhibit non-uniform DiF in its discrimination parameter is anxiety, found to be slightly stronger related to the wellbeing of older than younger adults.

Figure 31 – Density plot of predicted wellbeing from the ONS-4

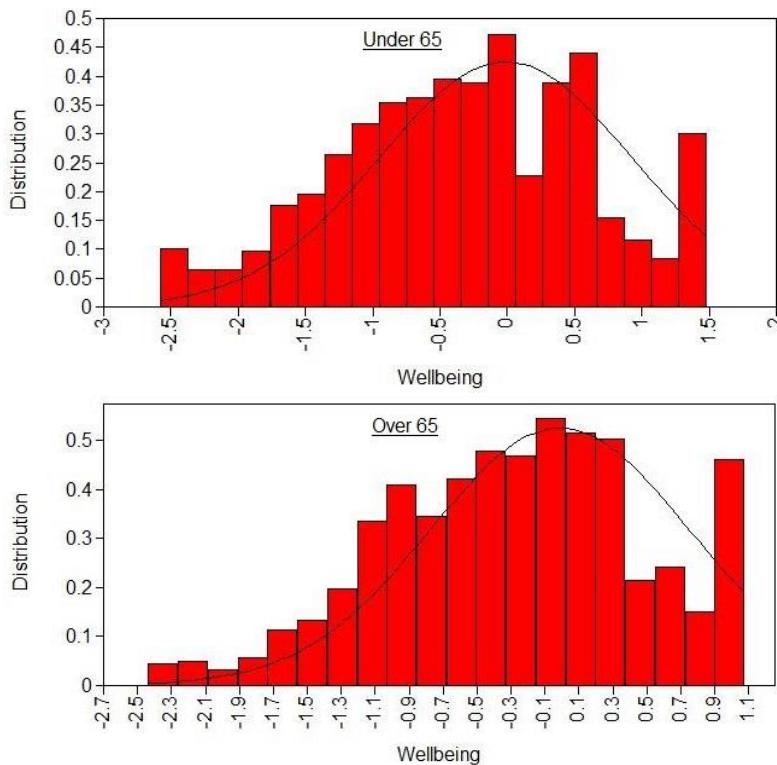


Table 26 – ONS-4 factor structures, unstandardised discrimination parameters and absolute model fit statistics

	Unstandardised discrimination parameters (SEs)	
	Under 65s	Over 65s
ONS-4	Wellbeing	Wellbeing
1. Life satisfaction	2.51 (0.06)	2.51 (0.06)
2. Worthwhile	2.18 (0.06)	2.18 (0.06)
3. Happiness	3.02 (0.09)	3.02 (0.09)
4. Anxiety	1.02 (0.04)	1.19 (0.06)
Factor mean	0	0.003 (0.04)
Factor variance	1	0.668 (0.04)
Model Fit	RMSEA mean (90% CI)	CFI
	0.055 (0.044, 0.066)	0.999

Discrimination parameters in bold exhibit non-uniform DiF as they were found to differ between under and over 65s

The difficulty parameters (Table 27), represent the amount of wellbeing required to have a 50% probability of responding above a certain category, signalling higher wellbeing. Anxiety requires the least amount of wellbeing to move between each of the categories signalling that people are more likely to respond higher to this question than the others. The b9 difficulty parameter for anxiety is low, at 0.494 for under 65s and 0.216 for over 65s, suggesting that anyone above these levels will respond at the ceiling of this item.

Difficulty parameters for happiness exhibited uniform DiF, as they were all higher in under 65s, meaning they required a higher level of wellbeing than older adults to be more likely to respond in a higher category. This means older adults are more likely than younger adults to respond higher and signal no problems with happiness. This was also the case for difficulty parameters b7-9 for the remaining three items, with older adults always more likely than younger adults to respond higher in the top few categories.

Constraining the unstandardized residual variances of both groups to 1 did not significantly impact the fit of the model ($p\text{-value}=0.159$) indicating that the amount of item variance not accounted for by the factor was the same across groups. The factor mean differed slightly between the two age groups, as shown in Table 26. The mean level of wellbeing is 0.003 SDs higher in the over 65 group, however this difference

was insignificant ($p\text{-value}=0.935$). The factor variance was found to differ between groups. As shown in Table 26, the over 65 group were less variable in wellbeing with a factor variance of 0.668, compared to the constrained factor variance of 1 in under 65s. Factor invariance was then tested to examine whether constraining the factor variance to equal 1 in both groups significantly impacted the model. The DIFFTEST was significant ($p\text{-value}<0.000$), suggesting that constraining the factor variance to 1 in both groups did significantly impact model fit and group factor variances do differ significantly.

Table 27 – ONS-4 difficulty parameters

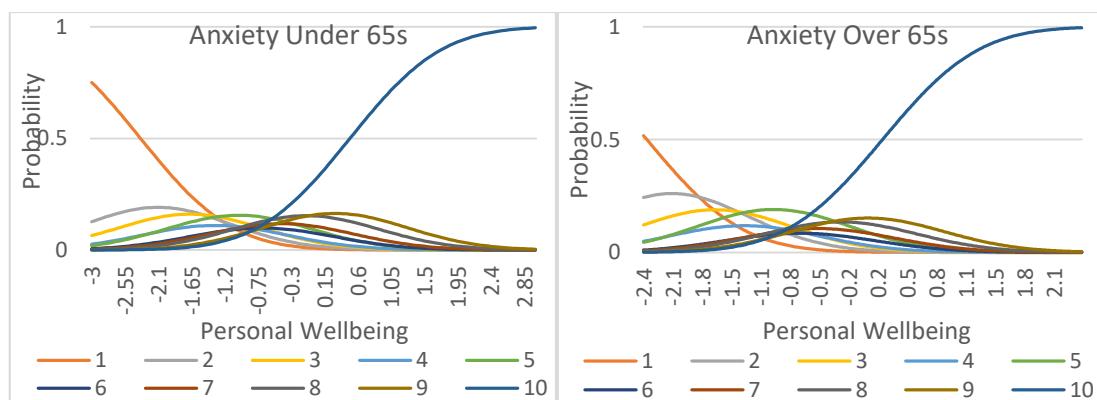
Difficulty Parameters (SEs)									
Under 65s				Difficulty	Over 65s				
life sat	worth	happy	anxiety		life sat	worth	happy	anxiety	
-1.68 (0.10)	-1.84 (0.10)	-1.73 (0.15)	-2.34 (0.07)	b1	-1.68 (0.10)	-1.84 (0.10)	-1.77 (0.21)	-2.41 (0.11)	
-1.39 (0.09)	-1.50 (0.09)	-1.42 (0.13)	-1.87 (0.06)	b2	-1.39 (0.09)	-1.50 (0.09)	-1.45 (0.16)	-1.85 (0.08)	
-1.08 (0.09)	-1.23 (0.08)	-1.18 (0.12)	-1.47 (0.05)	b3	-1.08 (0.09)	-1.23 (0.08)	-1.18 (0.14)	-1.45 (0.07)	
-0.86 (0.08)	-1.02 (0.07)	-0.92 (0.11)	-1.20 (0.05)	b4	-0.86 (0.08)	-1.02 (0.07)	-0.97 (0.13)	-1.20 (0.07)	
-0.52 (0.08)	-0.70 (0.07)	-0.62 (0.10)	-0.82 (0.05)	b5	-0.46 (0.09)	-0.70 (0.07)	-0.65 (0.12)	-0.80 (0.06)	
-0.26 (0.08)	-0.49 (0.07)	-0.39 (0.10)	-0.58 (0.04)	b6	-0.26 (0.08)	-0.49 (0.07)	-0.49 (0.11)	-0.62 (0.06)	
0.15 (0.08)	-0.11 (0.07)	0.00 (0.09)	-0.29 (0.04)	b7	0.05 (0.09)	-0.23 (0.08)	-0.17 (0.11)	-0.39 (0.06)	
0.81 (0.09)	0.48 (0.07)	0.57 (0.10)	0.09 (0.04)	b8	0.57 (0.10)	0.29 (0.09)	0.27 (0.12)	-0.11 (0.06)	
1.46 (0.11)	1.12 (0.08)	1.27 (0.13)	0.49 (0.04)	b9	1.08 (0.12)	0.77 (0.09)	0.75 (0.13)	0.22 (0.06)	

Where b1 represents the amount of QoL required to have a 50% probability of responding in the category signalling the lowest level of QoL and 50% chance of responding higher. Difficulty parameters in bold exhibit uniform DiF as they were found to differ between under and over 65s. Sat=Satisfaction Worth=Worthwhile

The ICCs for each item in under and over 65s are shown in Appendix 16. There are some issues in the performance of item levels. Categories 4 and 6 (responses 6 and 4 before the categories were reverse coded for this analysis) are never the most likely

option for any item in either group except happiness in under 65s where category 4 (response 6 from the original ONS-4 scale) is briefly the most likely choice. This may be due to the fact that people are drawn to 5 as the centre of the scale when they are trying to indicate a somewhat mid-level response. Response options 2, 3 and 7 (responses 8, 7 and 3 from the original ONS-4 scale) often also have smaller ranges over which they are the most likely response than more popular 1, 5, 8, 9 and 10 (responses 9 and 10, 5, 2, 1 and 0 from the original ONS-4 scale). These issues suggest that there are too many response categories and people may be struggling to distinguish some of them. These issues become extreme in the anxiety item (Figure 32) where, in younger adults only categories 1, 5 and 10 (responses 9 and 10, 5, and 0 from the original ONS-4 scale) have any range over which they are the most likely response. This indicates that for this item there is a strong tendency in younger adults to respond either that they are not at all, completely or moderately anxious in the centre. Categories 2 and 3 (responses 7 and 8 from the original ONS-4 scale) also have small ranges over which they are the most likely in older adults, but the ICCs still suggest substantial issues with levels for this item.

Figure 32 – Examples of problematic item characteristic curves from the ONS-4



In both age groups happiness provided the highest level of information, followed by life satisfaction, worthwhile and anxiety (Figure 33). The ONS-4 measure provided similar levels of total information in both groups across the lower end of the wellbeing scale (Figure 34), until approximately 1 SD above mean wellbeing where information drops quicker in over 65s than in under 65s. Internal consistency was good, with total information above 5, from approximately 3 SDs below the mean in both groups, up to

2 SDs above the mean in under 65s and approximately 1.6 SDs above the mean in over 65s.

Figure 33 – ONS-4 item information by age group

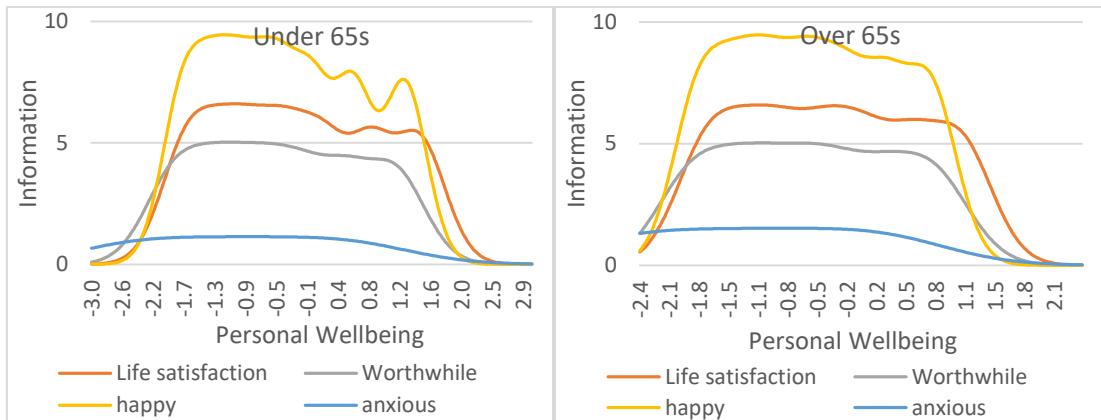


Figure 34 – ONS-4 total information by age group

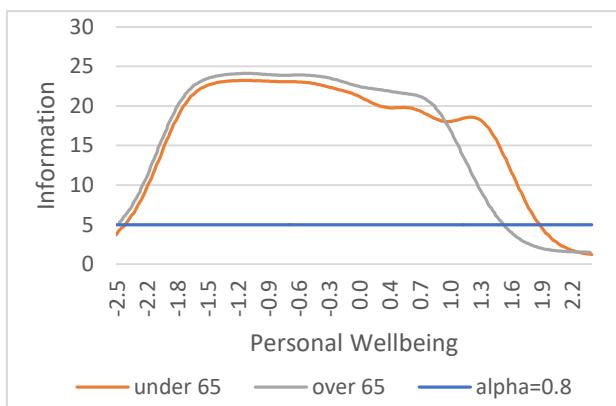


Table 28 – ONS-4 Item level DiF effect size

Item	Life Satisfaction	Worthwhile	Happy	Anxious	Total
ESSD	0.122	0.192	0.334	0.271	0.47

Older adults were expected to respond slightly higher to all questions (shown in Appendices 17 and 18). DIF had the largest impact on happiness and anxiety, which both had small effect sizes (Table 28). The effect size for worthwhile only just missed the threshold to be considered small (0.192 vs 0.2 threshold), indicating this may also have a small practical impact. The measure sum score had an ESSD of 0.47, indicating a small, but approaching moderate effect size. The maximum total impact of DiF (Figure 35) was 1.84 points difference in those 1 SD above mean wellbeing,

with older adults expected to respond higher. This is equivalent to 4.6% of the total score range (possible score range = 40) shown in Figure 36.

Figure 35 – Impact of DiF on ONS-4 Score

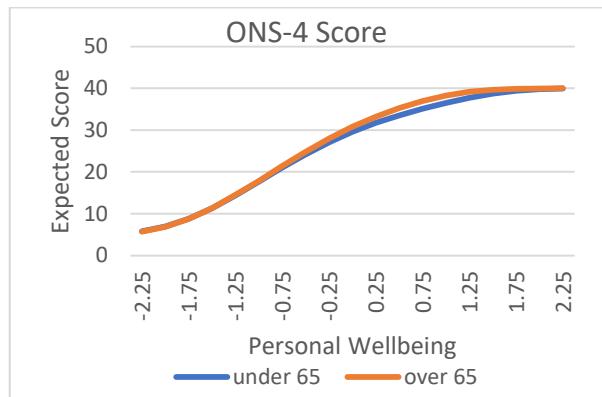
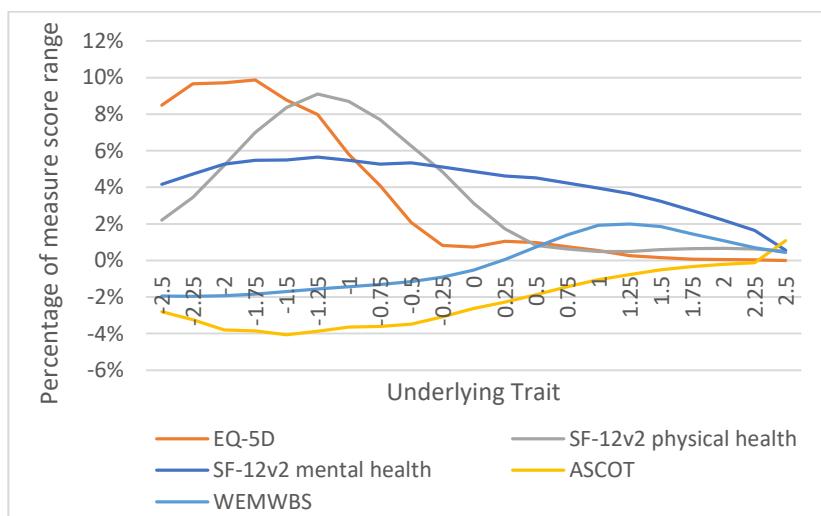


Figure 36 - DiF impact for each measure as a percentage of the potential score range



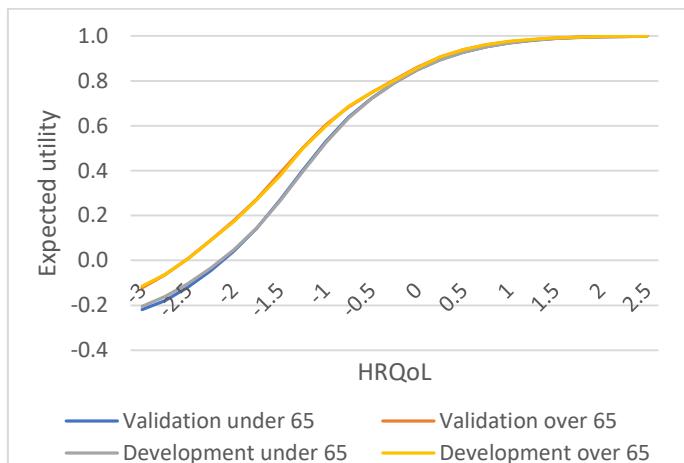
4.4.5 Validation

The final DiF model for each measure (obtained from the development sample) was rerun in the validation sample. Item parameters, expected scores and DiF impact were compared across samples to examine the stability of the DiF results found in the development sample. Stable results across samples can increase confidence in the results, as when DiF is identified in a large number of parameters, anchor parameters can be driven strongly by the sample, which can lead to volatile results regarding DiF direction and magnitude in different samples. Therefore, similar findings across samples can provide confidence in the robustness of DiF findings.

EQ-5D-5L

The final DiF model obtained in the development sample was rerun in the validation sample. Item parameters obtained from each of the samples were very similar, as shown in Appendix 19. Expected item decrements for each age group in the validation and development samples are compared in the figures displayed in Appendix 20. These figures show that DiF has a very similar impact on expected item decrements in both samples. The expected utility scores in the validation sample follow the same pattern as in the development sample (Figure 37), with over 65s expected to score higher than a younger adult with the same underlying level of health, until both groups start to reach the ceiling of the measure. In the validation set the difference in expected scores reaches a maximum of 10.5% of the utility score range of the EQ-5D-5L 2 SDs below the mean level of QoL. This is very similar to the maximum difference of 9.75% in the development sample, which occurred 1.75 SDs below the mean level of HRQoL. The expected scores of each group in the validation sample correspond almost exactly to those in the development sample, to the extent that the curves cannot be easily be distinguished from one another in Figure 37, suggesting that the results of the development sample are robust.

Figure 37 – Expected EQ-5D-5L utilities by age group in the development and validation samples



The DiF ESSD effect size results (Table 29) were also similar across the two samples. Both samples classified the DiF resulting from the first three EQ-5D-5L items as trivial. There were differences in the classification of pain/discomfort and anxiety/depression, which were both classified as small in the development set but classified as trivial and moderate respectively in the validation sample. However, this is likely due to the fact that the ESSDs of these items are estimated near the

cut-offs of the effect size classification system, as the point estimates of each item did not vary hugely between samples. In the development set the pain/discomfort item only just made the 0.2 cut-off for a small DiF effect size, while the anxiety/depression estimate was just below the 0.5 cut-off for moderate DiF. In the validation dataset there was some variation in the ESSD estimates which led to estimates for these items falling just the other side of the relevant cut-off line. The impact of DiF on the total EQ-5D-5L utility remained small in both samples.

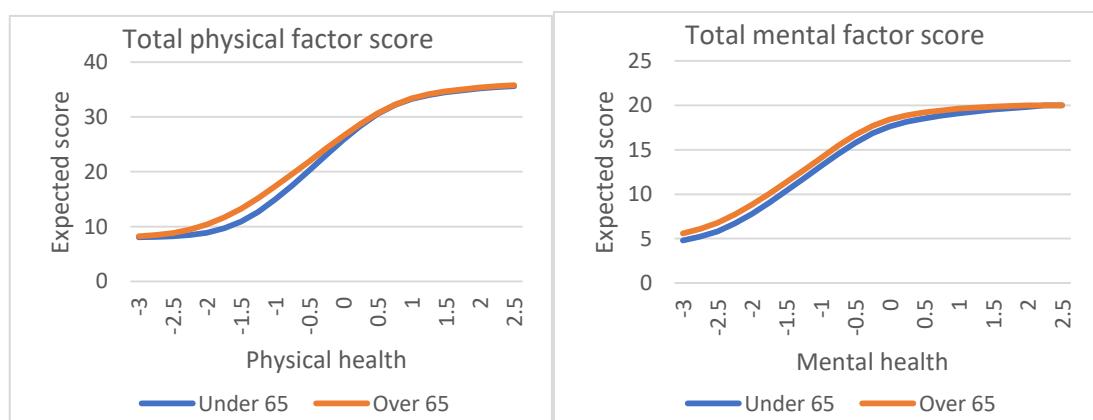
Table 29 – EQ-5D-5L DiF ESSD effect sizes in the validation and development samples

Item	Mobility	Self-care	Usual Activities	Pain/discomfort	Anxiety/depression	Total
ESSD development sample	-0.168	0.015	0.059	0.203	0.475	0.209
ESSD validation sample	-0.177	0.022	0.047	0.167	0.550	0.215

SF-12v2 TLA

Differences in item parameters produced from the final DiF model in the development and validation sample were minimal (Appendix 19). Expected item scores match well across the development and validation samples, as shown in Appendix 20. Total expected factor scores, shown in Figure 38, also correspond closely across the samples. The maximum difference between the expected physical and mental factor scores between age groups was 9.1% and 5.7% respectively in the development sample and 8.9% and 5.9% in the validation sample at 1.25 SDs below the mean level of health in all cases.

Figure 38 - Expected SF-12v2 TLA scores by age group in the development and validation samples



The effect size classifications of the impact of DiF on each item and total factor scores were similar between the two samples, as shown in Table 30. The only difference was the effect size classification of the total physical factor score. In the development sample, this was just over the moderate cut-off of 0.5, while in the validation sample the effect size did not reach this cut-off and the DiF was categorised as small.

Table 30 – SF-12v2 TLA DiF ESSD effect sizes in the validation and development samples

Item	General health	Physical Functioning	Physical role	Emotional role	Pain
ESSD Development sample	0.154	-0.132	0.233	0.000	0.338
ESSD Validation sample	0.137	-0.169	0.226	0.000	0.287
Item	Mental Health	Energy	Social activities	Total physical	Total Mental
ESSD Development sample	0.715	0.231	0.444	0.505	0.436
ESSD Validation sample	0.767	0.224	0.437	0.457	0.495

ASCOT

Item parameters and expected item scores corresponded very closely across the development and validation samples, as can be seen in Appendices 19 and 20 respectively. The total expected ASCOT score for each age group at any given underlying level of SCRQoL was almost equal across samples, as shown in Figure 39. The impact of DiF on the ASCOT utility reached a maximum of 4.1% of the score range in individuals 1.5 SDs below the mean SCRQoL level in the development sample and 5.0% of the score range in individuals 1.75 SDs below the mean SCRQoL level in the validation sample, with older adults expected to score slightly lower. The effect size classifications of the impact of DiF on each item and on the total ASCOT score was the same across samples, with ESSD estimates well matched across the development and validation sets (Table 31).

Figure 39 - Expected ASCOT scores by age group in the development and validation samples

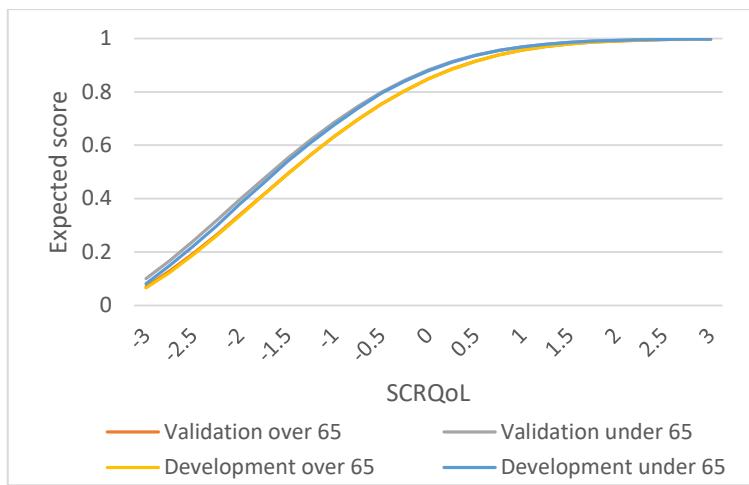


Table 31 – ASCOT DiF ESSD effect sizes in the validation and development samples

Item	Control	Personal clean	Food/Drink	Safety	Social	Occupation	Accommodation	Dignity	Total
ESSD Development sample	0.206	0.14	0.043	-0.068	0.07	0.394	0.009	0.166	0.37
ESSD Validation sample	0.217	0.184	0.069	-0.050	0.059	0.407	0.018	0.174	0.410

WEMWBS

Again, it can be seen in Appendix 19 that item parameters were fairly equal across samples. The item expected scores in Appendix 20 and the total expected WEMWBS score in Figure 40, demonstrate that DIF results obtained from the development sample were matched closely by results obtained from the validation sample. Total DiF impact remained below 2% of the possible score range (56) across the entire range of wellbeing tested in the development sample and reached a maximum of 2.5% in the validation sample, 2 SDs below the mean. ESSD estimates were also well matched across the development and validation samples, resulting in identical classifications of the effect size of DiF on item and factor scores, as shown in Table 32.

Figure 40 - Expected WEMWBS scores by age group in the development and validation samples

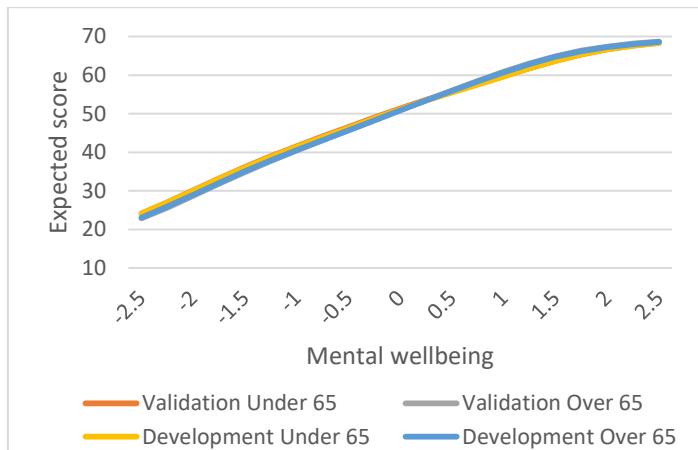


Table 32 - WEMWBS DiF ESSD effect sizes in the validation and development samples

Item	Optimist future	Useful	Relax	Interest people	Energy	Dealing problem	Think clearly	Feel good
ESSD Development sample	-0.345	-0.128	0.278	0.121	-0.239	0.072	0.119	-0.024
ESSD Validation sample	-0.350	-0.142	0.283	0.085	-0.299	0.076	0.115	-0.049
Item	Close people	Confid- ent	Make mind	Loved	Interest things	Cheerful	Total Internal	Total External
ESSD Development sample	0.05	0.014	0.08	-0.002	-0.162	0.044	0.12	-0.192
ESSD Validation sample	0.088	0.013	0.078	0.000	-0.167	0.066	0.093	-0.218

ONS-4

Item IRT parameters were consistent across the development and validation samples, as shown in Appendix 19. Expected item scores obtained for each age group from the validation and development samples are displayed in the figures in Appendix 20. These figures show that expected scores correspond well across the samples, with only small differences noted for the anxiety item.

The maximum difference in expected scores between under and over 65s was 3.58% of the possible score range in those 1.25 SDs below the mean level of wellbeing in

the validation sample, with older adults expected to score higher. This is compared to a maximum of 4.6% of the possible score range in those 1 SD above the mean level of wellbeing in the development set.

More variation was seen in the point estimates of the ESSDs between the development and validation samples for each item and the total score in this measure than the others (Table 33). However, the resulting classification of effect sizes remained the same, suggesting that the results surrounding the practical importance of DiF were consistent across samples.

Figure 41 - Expected ONS-4 scores by age group in the development and validation samples

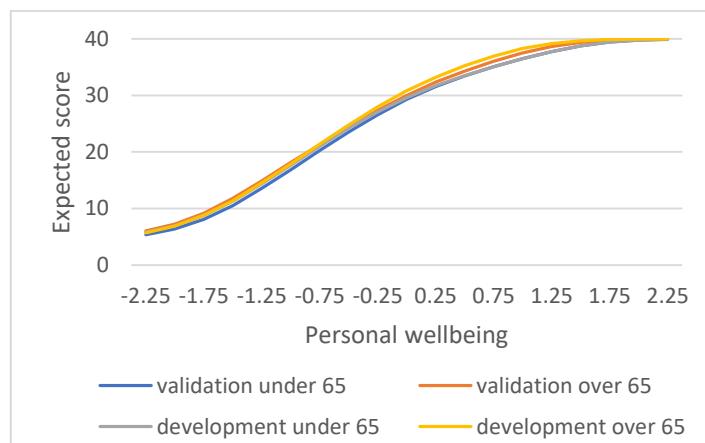


Table 33 – ONS-4 DiF ESSD effect sizes in the validation and development samples

Item	Life Satisfaction	Worthwhile	Happy	Anxious	Total
ESSD Development sample	0.122	0.192	0.334	0.271	0.47
ESSD Validation sample	0.059	0.045	0.252	0.393	0.373

4.4.6 Sensitivity analyses

4.4.6.1 Amended age cut-off – 75 years old

Results remained broadly similar using the higher age cut-off for classifying individuals as older adults. This higher cut-off could not be tested for the ASCOT as this dataset only provided information on whether individuals were under or over 65, without providing their specific age. Further details on any differences in results between the models using the different age cut-off are outlined below. Model parameters for the 75-cut-off model for each measure can be found in Appendix 21.

Patterns in DiF in discrimination and difficulty parameters for the EQ-5D-5L in the 65-cut-off and 75-cut-off models were broadly similar. The ranking of discrimination parameters was same across models. As in the 65-cut-off model, the discrimination parameters for self-care and pain/discomfort were lower in older adults. DiF impact was slightly higher in the 75-cut-off model across the range of traits (max 10.9% vs 9.7%).

More differences were seen in the DIF findings for the SF-12v2. The emotional role items remained DiF free across both models. In the 65-cut-off model non-uniform DiF was indicated with lower discrimination parameters in over 65s for the physical role items, pain, downhearted/low and social activities. However, in the 75-cut-off model non-uniform DiF was less widespread, with discrimination parameters lower in over 75s for physical role accomplish and pain but higher for energy. The pattern of DiF in difficulty parameters was very similar across models for the emotional role items, pain, calm/peaceful, energy, downhearted/low and social activities, however the direction of DiF in difficulty parameters was more mixed for general health, moderate activities, stairs and the physical role items in the 75-cut-off model. The impact of DiF was lower in the 75-cut-off model across the range of traits (max 4.7% vs 12.1%).

Uniform DiF in difficulty parameters of the WEMWBS follow similar patterns across the different age cut-off models. Discriminations for feeling optimistic about the future, useful, interested other people and feeling good about oneself required freeing due to non-uniform DiF, in 75-cut-off model while feeling useful, interested other people and able to make up own mind are freed in 65-cut-off model. In the 75-cut-off model, the discriminations for feeling optimistic about the future and feeling good about oneself are lower in over 75s while feeling useful and interested in other people are higher in over 75s. DiF impact was similar across models (max 2.8% vs 2.0% in 75 vs 65 model).

The anxiety discrimination of the ONS-4, which needed freeing in 65-cut-off model, remained the same across age groups in 75-cut-off model, meaning there was no non-uniform DiF in the 75-cut-off model. Difficulty parameters were all slightly lower in both groups but differences in difficulty parameters between age groups remained similar for all items except anxiety, for which the differences between groups are larger in the over 75-cut-off model. In the 75-cut-off model all difficulties for happiness and anxiety were lower for over 75s than under 75s. The impact of DiF is bigger in the 75-cut-off model than in the 65-cut-off model in below average individuals, but similar in above average individuals. The maximum impact is very similar across models (5.0% vs 4.6%).

4.4.6.2 Testing factors using combined EQ-5D SF-12v2 ONS-4 Model

A combined model including the EQ-5D-5L, ONS-4 and SF-12v2 was run to test whether forcing single factor solutions on the EQ-5D-5L and ONS-4 had an impact on the observed performance of items which may have covered slightly different concepts (such as anxiety/depression in the physical functioning focused EQ-5D and anxiety in the otherwise positively worded ONS-4). The combined model including these measures and the SF-12v2 was run to test whether the results regarding the factor structure and item performance of these measures obtained in the previous sections held.

The factor structure of the combined EQ-5D-5L SF-12v2 ONS-4 model was tested using EFA. The EFA eigenvalues (13.7, 2.3, 0.7) and scree-plot (Figure 42) suggested that two factors were present, with items split across them as shown below in Table 29. The first factor was a physical health factor containing all the items from the physical health factor in the SF-12v2-only model identified earlier in this chapter, plus the first four items from the EQ-5D-5L. The second factor was a mental health factor containing the items from the mental factor in the SF-12v2-only model earlier in the chapter, plus all four ONS-4 items and the EQ-5D-5L anxiety/depression item. Therefore, the only item from these three measures to move away from its original factor in its measure specific model was the EQ-5D anxiety/depression item. This could be anticipated as this item reflects concepts covered in the SF-12v2 emotional role and downhearted/low items and the ONS-4 anxiety item, which are all represented by the mental health factor.

Figure 42 Scree plot for combined model

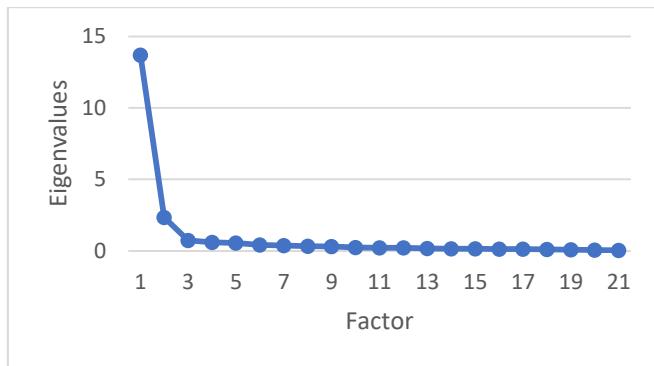


Table 34 – Factor structure of the combined model with discrimination parameters compared to the relevant single measure model

	Under 65		Over 65	
	Combined	Single	Combined	Single
Physical health				
EQ-5D Mobility	2.075	2.64	2.075	2.64
EQ-5D Self-care	2.049	2.469	1.652	1.858
EQ-5D Usual Activities	2.214	2.322	2.214	2.322
EQ-5D Pain/discomfort	1.728	1.685	1.728	1.299
SF-12 General Health	1.48	1.40	1.727	1.4
SF-12 Moderate activities	2.107	2.083	2.548	2.083
SF-12 Stairs	1.813	1.605	1.813	1.605
SF-12 Physical role accomplish	3.361	3.85	2.36	2.216
SF-12 Physical role limited	3.733	4.203	2.447	2.294
SF-12 Pain	1.883	1.62	1.883	1.226
SF-12 Energy	1.198	1.214	1.512	1.214
SF-12 Social activities	1.984	1.897	1.984	1.58
Mental Health				
EQ-5D Anxiety/depression	1.71	0.857	1.71	0.857
SF-12 Emotional role carefully	3.1	3.613	3.1	3.613
SF-12 Emotional role limited	3.331	3.886	3.331	3.886
SF-12 Calm/peaceful	1.089	1.089	1.337	1.089
SF-12 Downhearted/low	1.128	1.073	1.128	0.898
ONS-4 Life Satisfaction	2.351	2.511	2.351	2.511
ONS-4 Worthwhile	1.721	2.176	1.721	2.176
ONS-4 Happy	2.425	3.017	2.048	3.017
ONS-4 Anxious	1.18	1.024	1.18	1.188

The fact that the EQ-5D-5L anxiety/depression item moved away from the other four EQ-5D-5L items may mean that when a single factor solution was forced on the EQ-5D-5L, this item may have appeared to perform worse than it would in a more suited

factor. Some of its poor performance may have been due to it not fitting well in the unidimensional EQ-5D-5L model, despite the fact that the EQ-5D-5L only model had good fit statistics throughout the IRT and DiF process. The performance of this item may require further checking using methods which do not rely on the unidimensionality of the EQ-5D-5L scale to examine the extent to which the poor performance of this item is due to it not fitting with the dominant of the physical health factor in the EQ-5D-5L.

4.5 Discussion

4.5.1 Strengths/key findings

The results from this chapter provide important information about the psychometric performance of the included PROMs in measuring the health, QoL and wellbeing of older adults as well as younger adults. All included measures reported some problems. The EQ-5D-5L and ASCOT exhibited substantial ceiling effects for above average respondents in both age groups, resulting in reduced internal reliability and ability to discriminate the QoL of these respondents. The WEMWBS was able to discriminate the wellbeing of respondents across a broad range of wellbeing and had the largest range over which it achieved a good level of internally consistency, followed by the ONS-4 and SF-12v2, which were also internally consistent over broad ranges of their respective underlying traits in both age groups.

There was strong suggestion of item redundancy within the SF-12v2 mutli-item scales, resulting in a TLA super item approach being taken for this measure. There was also possible suggestion of item redundancy in the WEMWBS. There were occasional issues with the use of some response options across all the measures, but they are more widespread in certain measures. In the ONS-4 the eleven response options available did not appear to be used evenly. Respondents appeared to be drawn to either end of the scale and five in the centre. This suggests that there are simply too many to choose from and that they may not be being used as a smooth scale as intended. In the SF-12v2 TLA analysis, even scores dominated the ICCs of the super items, suggesting that there is a strong tendency for respondents to choose the same response option for each item within a super item pair, resulting in even scores.

DiF with at least a small effect size was found for a variety of items across all five measures. However, the issue was more widespread and had a stronger impact on expected item and measure scores in some measures than others. A particularly important finding is the presence of substantial DiF in the SF-12v2 and EQ-5D-5L. The impact of this DiF is particularly strong in respondents with below average health. This could be an issue when using results from these measures to evaluate interventions and make resource allocation decisions. Bias in scores of different age groups could affect decision making in many different ways. Within an evaluation for an intervention aimed at a broad age range of patients, it could cause different age groups to receive inappropriately different estimates of effectiveness. If subgroup analysis is conducted, this could result in the intervention only being provided to some individuals within the patient population while others are denied the intervention (which should have been cost-effective), based solely on their age. Conversely, an intervention could also be inappropriately approved in a subgroup in which it is not truly cost-effective, leading to a waste of resources. If the intervention is only aimed at a single age group, the effectiveness estimates could simply be lower or higher than they should in fact be, potentially leading to similar errors in decision making. At the NHS level interventions, which may only be appropriate for different age groups compete for funding. Therefore, bias in effectiveness estimates could unfairly bias funding decisions for or against certain age groups.

As the EQ-5D is currently the measure used by NICE to assess the incremental effectiveness and cost-effectiveness of new treatments, it was important to assess the impact that DiF within this measure may be having on these assessments. Hypothetical trial scenarios conducted for the EQ-5D-5L revealed that DiF impacts the estimates of effectiveness and incremental effectiveness that different age groups receive within a trial, which will in turn affect the ICER. The direction and size of this bias is not consistent and is dependent on individuals' position on the underlying trait. For conditions with extreme burdens of disease, with EQ-5D utilities worse than dead, older adults will likely receive higher estimates of effectiveness and therefore results will be in favour of interventions aimed at older adults. However, for conditions with baseline EQ-5D utilities above 0, it is likely that DiF would cause bias against older adults. This highlights the importance of understanding the impact of DiF and controlling for it within evaluations.

In terms of the aspects of measurement performance covered in this study, the WEMWBS appears to be the most appropriate measure for use in the evaluation of

health and social care services aimed at older adults. It is internally reliable over the broadest range of underlying trait; its response options perform well, and it exhibits the lowest level of DiF.

Previous studies have also identified age related DiF in items from the EQ-5D, SF-12 and WEMWBS. The investigation of DiF in the EQ-5D-5L in this study found that older adults with the same level of health as younger people were more likely to respond higher (signalling better health) to pain/discomfort and anxiety/depression and slightly lower to mobility. Similar results were found in a study of DiF in cancer patients using a Rasch Partial Credit Model (Smith, Cocks et al., 2016), which found age related DiF for pain/discomfort, anxiety/depression, mobility and self-care. Unfortunately, the direction of DiF effects was unclear in the reporting. DiF results for the SF-12v2 in this thesis found that older adults were more likely to respond higher (signalling better health) than a younger person with the same underlying level of health to a range of items including: general health, physical role, pain, the mental health item pair, energy and social activities. Some of these results were mirrored in other DiF studies, based on either the SF-12 (Fleishman and Lawrence, 2003) or the SF-36 (Lix, Wu et al., 2016, Teresi, Ocepek-Welikson et al., 2007, Yu, Yu et al., 2007). In this literature older respondents were also found more likely to respond higher to pain (Fleishman and Lawrence, 2003), calm/peaceful (Lix, Wu et al., 2016, Teresi, Ocepek-Welikson et al., 2007, Yu, Yu et al., 2007), energy (Fleishman and Lawrence, 2003) and downhearted/low (Fleishman and Lawrence, 2003). There is very limited existing evidence available investigating age related DiF in the WEMWBS. The only study available is a Rasch analysis of the WEMWBS, which aimed to identify a subset of items (which became the SWEMWBS), which formed a unidimensional scale and identified and eliminated redundant items (Stewart-Brown, Tennant et al., 2009). This study reported that some of the WEMWBS items exhibited DiF due to sex and/or age and were excluded from the SWEMWBS but did not report which items exhibited DiF due to age or in what direction. The only item that it did confirm exhibited age related DiF was feeling optimistic about the future, although the direction was not reported. However, this matches the fact that this item was found to be problematic in terms of DiF in this study.

An important strength of this work is that the psychometric analysis was conducted using an IRT and DiF framework rather than using CTT methods. As seen from the systematic review in Chapter 3, very little investigation of the psychometric performance of these measures in older adults has been conducted using IRT

methods. However, as detailed in section 4.3.1, IRT has several important methodological advantages over CTT methods, including much more detailed evidence on the performance of each item and estimation of internal consistency reliability and SEM which varies by trait level (rather than the single estimate of internal consistency provided by CTT). This provides important information on where along the underlying trait, and in which individuals, the measure is able to precisely discriminate the QoL of respondents.

The included measures differ in terms of what they aim to measure, and the specific concepts included, with EQ-5D and SF-12v2 focussing on health, ASCOT on the impact of social care on QoL, WEMWBS on mental wellbeing and ONS-4 on personal subjective wellbeing. These measures cannot be directly compared in terms of performance without additional qualitative consideration of what should be included in a broader QALY. It is important that the content and focus of this broader QALY aligns with the policy and service perspective which it is being used to evaluate, otherwise the impacts of these services will be missed, and they will continue to be undervalued and underfunded. While there are regular arguments for broadening the QALY beyond health, further work needs to be carried out to decide exactly what concepts are important to include in a comprehensive assessment of broader QoL and wellbeing and the full breadth of services which this broader QALY will be used to evaluate needs to be considered to be sure that the resulting measure is appropriate.

In addition to conceptual differences between the measures there are also methodological differences. An important consideration in the potential use of measures in the economic evaluation of health and social care is that they need to be preference-based, and that this needs to be on an appropriate scale for any broader QALY that results. Currently, the EQ-5D is preference-based on anchors of best and worst health imaginable using time trade off (TTO) exercises in the general population (Dolan, 1997), while ASCOT is preference-based on anchors of all to none of an individual's social care needs being met using best worst scaling exercises in social care users, anchored to death by a TTO exercise (Netten, Burge et al., 2012). The WEMWBS and ONS-4 are not currently preference-based, while the SF-12v2 is said to be preference-based using IRT methods (Maruish, 2012). Any future decision broadening the QALY may therefore involve not only a change of measure to one which comprehensively captures those aspects of QoL and wellbeing which have been found to be important to the broader QALY, but also an accompanying

preference elicitation using appropriate methods, in an appropriate sample using appropriate anchors for the resulting broader QALY.

When considering aspects of psychometric performance, it is also important to consider how the measures are administered. All measures included were completed on paper, so administration is comparable in this sense. However, the WEMWBS was completed on paper by the respondent during an extended face-to-face interview, in the presence of an interviewer, while all other measures were completed via postal questionnaire. While having participants complete the questionnaire within an interview may improve the proportion of participants who return at least a partial response and therefore improve data quality, it is not clear what impact this would have on item level response rates or on participants' responses. They may be more likely to respond more positively, if they feel the interview might read or judge their answers and they may feel more pressure to complete all items. However, it is also possible that completion during a lengthy interview increases burden on the respondent and they may disengage and rush, missing items. Therefore, the impact of the interview format on participant responses remains unclear but should be kept in mind during interpretation of psychometric results.

There are other aspects of measurement performance not covered in this work which also require consideration before a final choice on a preferred measure can be made. This phase of work covered areas of construct validity, internal consistency and the detailed measurement performance of individual items. The next chapter of this PhD details the methods used and results obtained from a qualitative study of the content validity of some of these measures in older adults. However further investigation is required into the remaining aspects of reliability (test-retest and inter-rater) as well as the responsiveness and sensitivity of candidate measures in a wide range of health and social care interventions and populations. These aspects require repeated measurements and therefore could not be picked up in this secondary data analysis. While the literature review was able to pick up on some existing evidence for these properties in older adults for the more established measures such as the SF-12 and EQ-5D, the measurement properties of the newer measures have yet to be extensively studied in specific populations such as older adults. Therefore, there is a need to continue gathering psychometric evidence on such measures in a wide range of health and social care populations to enable the selection of a best performing measure for the economic evaluation of both health and social care services.

4.5.2 Limitations

There are several limitations to this study which require discussion. Some of these limitations relate to the use of IRT methods to examine DiF. The accuracy of parameter estimation and DiF detection have been shown to be dependent on sample size, the type of IRT model used and model fit (Tay, Meade et al., 2015, Teresi, Ocepek-Welikson et al., 2009). Sample sizes of at least 500 per group have been recommended for stable parameter estimates (Tay, Meade et al., 2015). These were exceeded in all samples in this analysis, with the smallest group size being 940. A variety of different types IRT models were also fitted and the best fitting model chosen for each measure, which should minimise this issue. Model fit was judged to be good for all models according to the CFI and judged to be at least satisfactory ($\text{RMSEA} < 0.08$) for the EQ-5D-5L, ASCOT and ONS-4. The SF-12v2 and WEMWBS failed to meet the recommended RMSEA cut-off, suggesting there may be misfit and the potential for false DiF identification in these models. It has been suggested that the standard errors of item parameters should be checked as an indication of estimation accuracy, with a cut-off of $\text{SE} < 0.35$ indicating a good level of accuracy (Tay, Meade et al., 2015). This cut-off was achieved by all items in both measures. The similarity between results in the development and validation sets should also provide further confidence in the results obtained. Still, other methods of DiF detection should be tested on the same measures in future research to either confirm results found here or to examine the extent that misfit may have impacted these results.

Another potential disadvantage of IRT based methods is that they have high power to detect even very small differences in item functioning when samples are large (Meade, 2010). Large samples are recommended in IRT analysis and therefore the identification of statistically significant, but practically unimportant DiF, which in practice has a minimal impact on scores, is a risk. Effect size measures were estimated to assist in the interpretation of the impact of DiF findings, as recommended in the literature to reduce false DiF detection and over interpretation of the impact of practically meaningless DiF (Meade, 2010, Teresi, Ocepek-Welikson et al., 2009). Another advantage of this thesis is that it includes a cognitive qualitative study, which can provide additional evidence on the way older adults answer items on these measures. This can be helpful in either supporting or refuting these findings.

No single general population dataset could be found which included all measures of interest. It was not feasible within the resource and time constraints of the PhD to collect a dataset of the size required for stable IRT analyses which included all of the

measures of interest. Therefore, different UK datasets were used for each measure. These were carried out in different samples; one in the general population, one in people recently discharged from hospital and one in state funded social care users. This may limit comparability between measures as the samples are quite different. This is particularly true for comparison of ASCOT with the other measures as the sample of social care users may be particularly different to the other two. However, this sample is appropriate to the intended user of the ASCOT measures and is therefore appropriate.

As a measure of SCRQoL the ASCOT may not be appropriate to measure the QoL of the general population in the economic evaluation of both health and social care services. Substantial ceiling effects have already been shown to be an issue in this measure in a large sample of social care service users. This issue would be even more acute in the general population where the measure would likely have very little power to precisely discriminate the QoL of large proportions of the general population and would likely have reduced internal consistency.

4.6 Conclusion

This study provides important evidence on the structural and construct validity and internal consistency of several existing and commonly used generic measures of health QoL and wellbeing in older adults. IRT methods were adopted as they have rarely been used to assess the psychometric properties of these measures in this population. These methods provide rich and important evidence on structural and construct validity and internal consistency while overcoming some of the shortcomings of more commonly used CTT psychometric methods. The results of this study are important and relevant to the current debate of how to measure outcomes in the CUA economic evaluation of health and social care services. In the next phase of work in this thesis, qualitative techniques were used to assess the content validity of a selection of the measures assessed above in older adults to further contribute to understanding of the psychometric performance of these measures in an important population for the evaluation of health and social care services.

4.7 Glossary of terms

Term	Description
Akaike Information Criterion (AIC)	A measure of relative fit between models, with lower values signifying a better fitting model
Bayesian Information Criterion (BIC)	A measure of relative fit between models, with lower values signifying a better fitting model
Comparative fit index (CFI)	A measure of absolute model fit bounded between 0 and 1 with higher values signifying better model fit. A common cut-off for good model fit is $CFI \geq 0.95$
Confirmatory Factor Analysis (CFA)	A statistical technique used to examine the factor structure of a set of observed variables and test whether a relationship exists between these variables and the underlying latent construct they are supposed to be measuring
Constrained Graded Response Model (constrained GRM)	A one-parameter version of the GRM
Differential Item Functioning (DiF)	DiF occurs when an item functions differently between subgroups of respondents. Where DiF is present, respondents with the same level of QoL but who belong to different subgroups, have a different probability of providing the same level of response to the item
Difficulty parameter (b)	An IRT model parameter (an item with n response categories has n-1 difficulty parameters) which tells us the amount of underlying trait required to have a 50% probability of responding above a certain category, signifying higher levels of trait, and a 50% chance of responding in that category or below
Discrimination parameter	An IRT model parameter (one per item) which examines how closely an item is related to the underlying trait of respondents
Exploratory Factor Analysis (EFA)	A statistical technique used to identify the factor structure of the relationship between a group of variables and one or more latent traits
Generalised Partial Credit Model (GPCM)	A two-parameter IRT model extension of the PCM
Graded Response Model (GRM)	A two-parameter polytomous ordinal IRT model
Information	A measure of the precision of measurement and internal consistency reliability of an item or measure in item response theory
Item characteristic curve (ICC)	ICCs (one per response category) describe the relationship between an individual's level of underlying trait and their probability of responding in each possible response category for a single item
Local dependence (LD)	A violation of the assumption of local independence required to fit an IRT model. Local dependence

	arises when there is additional systematic covariance between items beyond their given relationship to the underlying trait being measured.
Measurement Invariance	A statistical property of measurement which states that the same underlying construct is being measured across groups
Modification Indices (MIs)	Part of the model output in MPlus which show sources of local misfit in the model and suggest changes which could be made to improve fit
Partial Credit Model (PCM)	A one-parameter polytomous ordinal Rasch family model
Rating Scale Model (RSM)	A one-parameter polytomous ordinal Rasch family model
Root mean square error of approximation (RMSEA)	A measure of absolute model fit bounded between 0 and 1 with lower values signifying better fit. A common cut-off for good model fit is $\text{RMSEA} \leq 0.05$ and acceptable model fit is $\text{RMSEA} \leq 0.08$
Standard error of measurement (SEM)	The standard deviation of error of measurement in a test
Unidimensionality	A unidimensional scale measures only one underlying trait, to which all test items are related

Chapter 5

Qualitative investigation into the content validity of currently used health and wellbeing measures in older adults

5.1 Introduction

As has been previously discussed, health and social care services for older adults aim to improve or maintain not only the health of older adults, but also their broader QoL including aspects such as independence and social participation (van Leeuwen, Bosmans et al., 2015b). There is concern that traditional measures of HRQoL may miss these broader benefits and therefore these services will be undervalued in economic evaluation. It is important that measures used to evaluate the impact of health and care services aimed at older adults are valid in assessing the QoL of older respondents and are acceptable to them.

Recent guidelines for measure development advise that content validity should be assessed in relevant groups of respondents during the development of measures to ensure that questions are understood, relevant, appropriate and that the measure is comprehensive in its coverage of important aspects of the construct being measured (Brod, Tesler et al., 2009). However, there is little evidence that patients or public were involved in the development of the EQ-5D-3L, SF-36, and resulting SF-12v2 or the ONS-4 through testing of content validity. For these measures, developers and experts generated domains and items and then testing was mainly quantitative. The face validity of the EQ-5D-5L was tested in members of the general population in the UK (which included eight individuals who were either retired or in receipt of a pension), but this was with the aim of testing the understanding of the new response levels and therefore the content validity of the questions themselves was not the focus (Herdman, Gudex et al., 2011). Some content validation was carried out in the ONS-4 questions once they had been released (which included eight participants aged 61+) (Ralph, Palmer et al., 2011). This work raised some issues with the ONS-4 and suggested some potential solutions. However, the questions proceeded unchanged. The WEMWBS was developed based on an existing measure, the Affectometer 2

(Tennant, Hiller et al., 2007). In the process of development, the research team carried out content validation interviews on the Affectometer 2, including two focus groups with older adults. Following content validation and statistical psychometric testing, they greatly reduced and altered this measure into what became the WEMWBS (Tennant, Hiller et al., 2007). At this point, they carried out two additional focus groups to check the content validity of the WEMWBS but did not include older adults. Therefore, the content validity of the WEMWBS in older adults was unknown. Due to the limited study of the content validity of the included measures in older adults this study represents an important contribution to knowledge.

In the previous chapter, IRT methods were used to examine the performance of five existing QoL and wellbeing measures in older respondents. Some issues with item response and DiF were found which present a threat to the construct validity of some of the measures. Qualitative methods of cognitive interviewing can be used to further explore issues identified using statistical psychometric methods by probing into the response process to identify response issues (Patrick, Burke et al., 2011b).

In addition to construct validity, it is also important that a measure had evidence of good content validity in the population in which it is being used. The questions and response options contained in the measure should be relevant to the QoL of older respondents and should comprehensively cover what is important to their QoL. The questions also need to be understood by respondents in the way developers intended and acceptable to respondents to ensure that respondents are willing and able to provide answers to the questions (van Leeuwen, Bosmans et al., 2015b).

In this chapter, qualitative methods will be used to examine the content validity of four of these measures in older adults. This will both further investigate issues found in the previous chapter and go further into seeking to find which measures are able to provide a valid and comprehensive estimate of what is important to the QoL of older adults.

5.2 Aims and objectives

To use cognitive interviews to examine the content validity of the EQ-5D-5L, SF-12v2, ONS-4 and WEMWBS in assessing the QoL and wellbeing of older adults.

5.3 Methods

5.3.1 Study design - Choice of data collection method

Qualitative methods have been widely used to examine issues with content validity (Buers, Triemstra et al., 2014, Collins, 2003). As seen in section 2.5.2.3, qualitative methods can and should be used at several points during the development and use of a PROM. The precise qualitative methods used may depend on which phase of PROM development or use we are at. Interviews or focus groups with respondents and experts can be used to generate domains and items which may be relevant to the construct of interest (Patrick, Burke et al., 2011a). Once a set of domains and items have been selected to form a PROM, the content validity of that measure should be checked using cognitive interviewing methods to examine how respondents react and respond to the measure in practice (Patrick, Burke et al., 2011b).

Cognitive interviewing methods are widely used to examine the content validity of existing PROMs (Rothrock, Kaiser et al., 2011) and are increasingly considered an essential aspect of instrument development because of the in depth evidence they can provide in support of content validity or the need for further instrument refinement (Knafl, Deatrick et al., 2007, Patrick, Burke et al., 2011b). Cognitive interviewing methods are broadly used and recommended for this purpose because, unlike quantitative methods, which can signal potential issues with items, such as high non-response rates, cognitive interviewing methods can go beyond this by examining the process that respondents go through when providing responses to questionnaires and identifying the causes of issues and appropriate solutions (Knafl, Deatrick et al., 2007). Cognitive interviews are able to test the assumption of shared understanding of items and concepts between measure developers and respondents (Patrick, Burke et al., 2011b) and examine the relevance and comprehensiveness of included domains to the concept of interest. This enables the maximisation of validity and reliability and the minimisation of measurement error of data obtained from PROM responses (Knafl, Deatrick et al., 2007, Patrick, Burke et al., 2011b). Since this study focuses exclusively on examining the content validity of existing PROMs, cognitive interviewing methods were chosen.

Cognitive interviewing techniques, based on theories of survey response, are often used to explore the process which respondents go through when answering survey questions (Collins, 2003). One of the most commonly seen theories of survey response is the question and answer model, developed by Tourangeau, which details

four stages that respondents go through when answering a survey question; comprehension, retrieval, judgement and response (Tourangeau, Rips et al., 2000), which was outlined in section 2.5.2.3. First, the respondent has to understand the question (comprehension), then they must retrieve the valid information from their memory (retrieval), make a judgement about the information needed to answer the question (judgement) and lastly, they must choose a response to the question (respond).

There are a variety of points in this process where issues may arise which threaten the content validity of the measure. Respondents may not understand the question or response options or may interpret them differently to how the measure developers intended. There may be a mismatch between the options provided and the desired response of the respondent. Or the respondent may provide an answer which is inconsistent to what would be expected, given what the respondent has said elsewhere in the interview or what the interviewer knows, or can see, about the respondent. Inconsistent responses can arise for several reasons. The respondent may feel that the questions asked are not relevant or appropriate to them, and therefore they may disengage and simply select an answer out of a sense of duty to respond, whether or not it applies to their situation. The respondent may feel social pressure to respond in a certain way. Or the respondent may have adapted to issues which have led to changes (most often declines) in health, QoL or wellbeing. These adaptive mechanisms are called response shift (Sprangers and Schwartz, 1999). Response shift involves a series of cognitive processes by which, in the face of declining health or functioning, respondents adapt and adjust their internal standards, values and conceptualisation of health; allowing them to continue to view their state as positive or stable (Spuling, Wolff et al., 2017). There are several types of response shift, which can impact individuals' answers in different ways (Sprangers and Schwartz, 1999). Response recalibration occurs when participants adjust their internal standard, or benchmark, of what they consider to be good health or QoL. For example, in response to a decline in mobility, respondents may lower their internal standards of what constitutes good mobility from being able to go for a long walk to being able to walk to a nearby supermarket and back. Response reprioritization occurs when participants reprioritize what is important to their health or QoL. For example, in the face of issues with physical functioning, respondents may place less importance on physical activities and more importance on mental and social elements of their health or QoL. Finally, response reconceptualization may occur, where participants not only reprioritise the relative importance of aspects of their health or

QoL, but they may change their definition of health or QoL such that an aspect which was previously important is no longer considered relevant, and vice versa. All these issues can limit the validity of data provided by PROMs and conclusions and comparisons based on responses provided.

Two commonly used cognitive interviewing techniques are think-aloud and verbal probing. Think-aloud techniques ask respondents to verbalise their thoughts as they complete a questionnaire. Verbal probing involves asking respondents specific questions in order to understand how they arrived at their chosen response either during the completion of a questionnaire (concurrent verbal probing) or after questionnaire completion (retrospective verbal probing) (Collins, 2003).

Studies in the literature using cognitive interviewing methods to investigate the content validity of measures of QoL and wellbeing have used either one-to-one interviews or focus groups. This has an impact on the type of cognitive interviewing method which can be used as think-aloud is only suitable for one-to-one interviewing while verbal probing can be used in either method. There is debate surrounding which method is more appropriate and effective for this type of study question. Arguments for both sides from the literature and from PPI for this study are discussed here.

In terms of the depth of individual's views obtained, one-to-one interviews are argued to provide a more in-depth insight into individual participant's views (Brod, Tesler et al., 2009). However, focus groups, by including more participants in any one data collection session, provide a wider range of views. The ability to get the views of a larger sample of respondents may also increase confidence in results. Focus groups also allow and encourage discussion amongst participants which can stimulate new views and ideas (Barbour, 2010). The questions posed by this topic may not be ones that people have ever contemplated at length. Therefore, group dynamics and discussion may stimulate additional thoughts and opinions which people may not have thought to express in a one-to-one setting (Brod, Tesler et al., 2009).

However, there are important potential disadvantages to focus groups which require consideration. Group thinking and dynamics is one important potential limitation. Discussion and views will be affected by the group (Barbour, 2010). There is a social tendency in groups towards agreement, which may result in individuals feeling they cannot easily express opposing views. The moderator must be mindful of this and provide opportunity for opposing views to be expressed if it is felt that these are not

being freely expressed in discussion. Disagreement may also lead to conflict, which also has to be managed.

One-to-one interviews are argued to be better than focus groups when topics are potentially sensitive, as discussion of aspects of QoL may be. People may be less willing to discuss sensitive topics in the presence of additional people who they do not know. However, it has also been argued that focus groups can in fact provide a safety in numbers for respondents (Barbour, 2010). There is not the pressure of one-to-one interviews, in which people feel obliged to answer every question. In a focus group, those that feel comfortable discussing a topic can answer it and those who do not can choose to stay quiet. Expression of honest and frank opinions from others can also encourage shier participants to express themselves freely. This can provide a greater level of control to participants. Brod et al argue that one-to-one interviews and focus groups should be viewed as complementary techniques and not either/or as they are separate and valid techniques which provide different information (Brod, Tesler et al., 2009).

From a practical point of view, there was concern that a substantial proportion of an older population experiencing frailty may be unable to travel to a suitable focus group location, even if a wheelchair accessible one was chosen. One-to-one home interviews are more appropriate in this hard to reach group which is often underrepresented in research, but who make up an important group of health and care service users. Therefore, home interviews may enable the participation of a wider range of older adults and make the sample more representative of the elderly population.

The choice of which type of data collection methods to use were debated within the research team as well as in early patient and public involvement (PPI) work. Members of the public were consulted at several stages throughout the study design and preparation phase. Early on in the design of this phase of the research, in April 2017, the researcher met with representatives of the ongoing Community Ageing Research 75+ (CARE 75+) study {National Institute for Health Research, 2014, The Community Ageing Research 75+ (CARE 75+) cohort study}. Since December 2014 the CARE 75+ study has developed a cohort of over 900 community dwelling older adults aged 75 and above. The aim of the CARE 75+ study is to investigate frailty transitions over time as well as collect health, social and economic data and act as a platform for additional studies aiming to improve outcomes for older adults. The aim of this

meeting was to discuss aim of the project and its potential design, in the hope that recruitment could be conducted through the cohort. This meeting included members of the CARE 75+ Frailty Oversight Group. The Frailty Oversight Group is an independent older lay reference group, comprising 8-10 members drawn from local stakeholder organisations such as the Bradford and District Older People's Alliance, The Older People's Advocacy Alliance Sheffield and Health Watch organisations across Yorkshire. This group have advised on the conduct of the CARE 75+ and related studies from the cohort's inception. Their approval was required for this study to recruit through the CARE 75+ cohort and their advice on design was thought to be important, since they know the cohort and the types of study they respond well to.

During this meeting, an overview of the project was provided by the researcher. The Frailty Oversight Group lay members were then asked whether they thought this project should be linked with the CARE 75+ study and whether they thought this was something the cohort would be interested in participating in. They agreed that this study would be of interest to the cohort and could be recruited through the CARE 75+ cohort. They were also asked to comment and give advice on the data collection methods. The Frailty Oversight Group suggested that providing potential participants the choice between one-to-one interviews and focus groups would maximise the comfort and control of participants and generate a wider range of data to consider. They also noted that isolation was a common issue in an elderly cohort and therefore, while some participants may find it more convenient to be interviewed in their own home, others may appreciate the opportunity to get out and meet a group of people. Therefore, based on PPI advice and arguments from the literature, it was decided to offer participants the choice between attending a focus group or one-to-one home interview, with each participant only required to attend one, not both.

However, although both forms of data collection were planned for and both options were provided to participants, during recruitment only two participants responded saying that they would be interested in a focus group. These responses arrived at very different times during the recruitment process and therefore it was not possible to run any focus groups. Both participants were happy to participate in one-to-one interviews instead and therefore only cognitive interviews, using a combination of think-aloud and retrospective verbal probing, were undertaken. A combination of think-aloud and verbal probing is recommended in the literature to maximise the amount of information gained on participants' interpretations and opinions of a questionnaire (Buers, Triemstra et al., 2014, Priede and Farrall, 2011). This allowed

for an in-depth exploration into the response process of individual respondents when completing each of the measures of interest.

Interview/focus group protocol/schedule

As discussed above due the preference for one-on-one interviews and not focus groups, data were collected solely through interview. These took place at a time and location suitable for the participants, mostly their own homes.

The home interviews were scheduled for 90 minutes, although completion of the topic guide was not anticipated to take this long. This allowed plenty of time for participants and the interviewer to chat at the beginning or end of the session and for a refreshment break in the middle of the interview if desired by the participant. Home interviews were adapted to the needs of the participants. If they felt it was too much, the interview could be spread over multiple visits or lengthy breaks taken. Interviews were audio recorded. The researcher also took field notes which allowed the incorporation of non-verbal cues such as participants expressions into the analysis.

To reduce participant burden each interview discussed two of the four measures of interest. All possible combinations were provided (as shown in Table 38), and an attempt was made to provide each combination to a similar number of participants. The researcher also attempted to balance the number of times each measure was discussed first within each combination of measures, in order to reduce the potential for interviewer-imposed bias.

A semi-structured topic guide, outlined in Table 35, was developed by following similar qualitative validations of QoL and wellbeing measures from the literature (Clarke, Friede et al., 2011, Milte, Walker et al., 2014, Taggart, Friede et al., 2013, van Leeuwen, Bosmans et al., 2015b) and recommendations for best practice guidelines for using qualitative methods for assessing content validity (Brod, Tesler et al., 2009). The interviewer began by introducing participants to the topic and how the session was going to run. The interviewer then asked participants to complete a brief demographic questionnaire (shown in Appendix 22) detailing their gender, age, level of education, ethnicity and a yes, no, question asking whether they have any long-term conditions. These variables are important as they are found to impact peoples'

understanding of questions and their responses in the literature (Fayers and Machin, 2016).

The interview then began with several background questions about the participant's life, family and living situation. The topic guide then moved into a discussion of the definitions of QoL and wellbeing and what is required in life to achieve a good level of these. Then the researcher explained the think aloud process and the first PROM was provided. Participants were asked to think aloud, saying whatever they were thinking while completing the measure and were prompted to continue thinking aloud if they became silent and stopped explaining how they were arriving at their answers. Once they had completed the measure they were asked for their initial impressions of the measure as a whole in terms of whether they found it clear, easy to understand and of acceptable length.

Verbal probing questions were then used to further explore participants' interpretation and understanding of terms in each question and whether they felt the questions were relevant and important to their QoL and wellbeing and acceptable to ask to someone like themselves. Once each item had been discussed, participants were asked whether they felt there was anything additional that was important to their QoL which had been missed from the measure, or whether they felt it gave a comprehensive view of their QoL or wellbeing. Then a break was offered, after which the second measure was provided, completed and discussed in the same manner. Participants were then given the opportunity to make any remaining comments about each of the measures before the topic guide closed by asking participants to indicate and discuss which of the two measures they preferred. Standardised questions were used throughout in order to maintain consistency across interviews and minimise the risk of interviewer induced bias, as recommended when conducting cognitive interviews (Willis, 2005).

Further PPI was carried out, focussed on finalising the topic guide and, study documentation and study details before commencing recruitment. This was conducted with an online advisory panel linked with Sheffield Teaching Hospitals which specialises in reviewing study documentation for researchers (Sheffield Teaching Hospitals NHS Foundation Trust, 2019). This advisory panel contained members of the public of various ages and backgrounds, with no single condition specific focus. Although the panel usually conduct their reviews online, they have one annual face-to-face meeting, at which this project was discussed in November 2017.

Table 35 – Interview topic guide

Introduction	Go through and discuss information sheet and consent form Check permission to record Give an outline of what will be done in the interview Fill in attribute questionnaire
Background questions	Warm up questions about the participant to get them comfortable such as: How long have you lived in this area/house? How is the area? Do you have family who live nearby?
Quality of life and wellbeing questions	Could you tell me what the term quality of life means to you? What do you feel you need in life to have a good quality of life? And wellbeing – what does that mean to you? Does it differ from quality of life?
Explain think aloud	Now I am going to give you the first questionnaire to fill out. I would like you to think aloud as you fill it out. What I mean by “think aloud” is that I would like you to tell me everything you are thinking from when you first see the question. You do not need to plan what you say or try to explain to me what you are saying. Just act as if you are alone speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time, I will ask you to keep speaking out loud. Do you understand?
Provide and complete 1 st questionnaire	
Verbal probing questions	How did you find that to complete? What does (term or item) mean to you? What did you think about when answering this question? Do you think this is something appropriate to ask someone like yourself? Do you feel this question is relevant to your quality of life? Do you think that there is anything important to your quality of life or wellbeing, which is missing from the questionnaire and should be included? Were there any questions which you felt were not relevant to quality of life or wellbeing, or not important for people like yourself?
Repeat think aloud and verbal probing with 2 nd questionnaire	
Measure preference	Now that we have discussed both questionnaires, was there one which you preferred? Did you think one of the two would do a better job of measuring your quality of life? Why is that?
Conclusion	Do you have any other comments you would like to make about anything we discussed today? Thank for participation

The topic guide, participant information sheet and invitation letter were provided to the panel for comment ahead of the meeting. At the meeting the background and rationale for the project were presented by the researcher before the panel were given the opportunity to comment on the study documentation and details. The panel requested some amendments to the wording of several aspects of the study documentation and topic guide to improve clarity. For example, they noted that it was not clear that members of focus groups also needed to maintain confidentiality as well as the researcher. They also suggested some additions to the topic guide including outlining that any questionnaires completed during the interview or focus group are solely for this study and would not be passed on to any care providers or the CARE 75+ study.

Pilot interviews

Two pilot interviews were carried out in January 2018, prior to starting data collection. These were generally successful, with pilot participants understanding the process of think aloud and verbal probing questions and not finding the topics upsetting. The pilots were also completed well within the estimated 90-minute interview time suggested in the participant information sheets, with each pilot being completed in between 45-60 minutes. The pilots each led to minor changes in the topic guide. It was felt after the first pilot that beginning the interview with questions about the meaning of QoL and what was needed to achieve a good QoL was difficult and the participant looked uncomfortable and struggled to provide full answers. Therefore, several background questions about the participant's life, previous work and living arrangements were added to allow them to become comfortable with answering questions before these more difficult QoL questions were asked. This worked much better in the second pilot.

In the second pilot the participant seemed generally comfortable and answered well however, at the end, during discussion about the interview it became clear that they had thought that these were questionnaires that the interviewer had made and that they had therefore been reluctant to be too negative about the questions within them. Therefore, extra background was added into the beginning of the topic guide to make it clear that these questionnaires had not been designed by the researcher and that the researcher was seeking people's honest opinions, positive or negative, about them.

5.3.2 Selection of measures

Consideration of participant burden was central to the design of this study. Cognitive interviewing methods are fairly demanding on the participant as they are required to describe their thought process out loud when answering each question and are then probed with further questions about how they arrived at their answer for each question. This is both time consuming and cognitively demanding. The measures investigated in this thesis vary substantially in length, from four to fourteen items. It would be impractical to ask participants to go through this process for all five measures. It was important to the study design that measures be discussed in all possible combinations in order to investigate their preferences between measures. Therefore, some participants would receive both the WEMWBS and SF-12v2, resulting in an in-depth discussion of 26 items in the content validation part of the interview. This was already considered a substantial burden. It was therefore decided that any one interview could only cover two instruments.

Ensuring that five measures were covered by a minimum of 10 participants (the minimum sample size required by the COSMIN checklist to consider a qualitative content validation study of excellent quality (Mokkink, Terwee et al., 2010)), would require at least 28 interviews, or some combination of fewer interviews and focus groups. However, it was unknown if this number would be sufficient to reach data saturation. Four measures could be discussed by 10 participants each, in 20 interviews, already substantially reducing the resource and time burden of this study, in a restricted PhD timeframe and budget. Therefore, it was felt important to reduce the pool of measures included in the qualitative study.

The systematic review presented in Chapter 3, revealed no evidence on the performance of the two wellbeing measures in older adults specifically. With wellbeing measures being mentioned as possibly appropriate for the evaluation of social care interventions (National Institute for Health and Care Excellence, 2016) it was felt that keeping the two wellbeing measures was a priority, especially given the local issues experienced when using these measures in older samples. The EQ-5D-5L was also felt important to keep, as this measure is the current standard practice in healthcare evaluation and is being claimed as potentially inappropriate for the evaluation of social care interventions. The SF-12v2 was originally added at the suggestion of the research group behind the CARE 75+ cohort study as they felt that the SF measures represented a more balanced view of the health and QoL of older adults by including questions about social contact, such as the social activities item. The two health

measures also exhibited substantial DiF in the quantitative psychometric validation study presented in Chapter 4. It was felt that qualitative investigation into the process of responding to these questions may provide more understanding as to why these issues were arising.

Part of the original aim of this work was to find a measure, which was suitable for the evaluation of social care interventions as well as health interventions, in line with increasing calls for integration between health and social care services, while maintaining comparability between evaluations by using the same measure across evaluations. ASCOT is a measure of SCRQoL, strongly focussed on those aspects of QoL which are impacted by social care services. It has large ceiling effects even in a social care population. It is therefore likely that these ceilings would be even higher if it were used to evaluate health interventions, particularly those with a low burden of illness which do not substantially limit patients in daily activities. This would limit its ability to detect change in QoL resulting from such interventions. Therefore, it was felt that the ASCOT was the least likely measure to be broadly appropriate in all populations and interventions across health and social care evaluations and this measure was excluded from this content validation study. This measure was also the measure which had incorporated the most input from older adults during measure development. This included cognitive interviews with social care users of various ages, which included approximately 15 individuals over the age of 65. Where issues arose, items were amended, and further cognitive testing was used to check their performance. Therefore, of the measures included in this study, this is the measure which has the most evidence in support of its content validity in older adults and therefore examining its content validity is of lower priority in comparison to the other measures.

5.3.3 Recruitment strategy and sample size

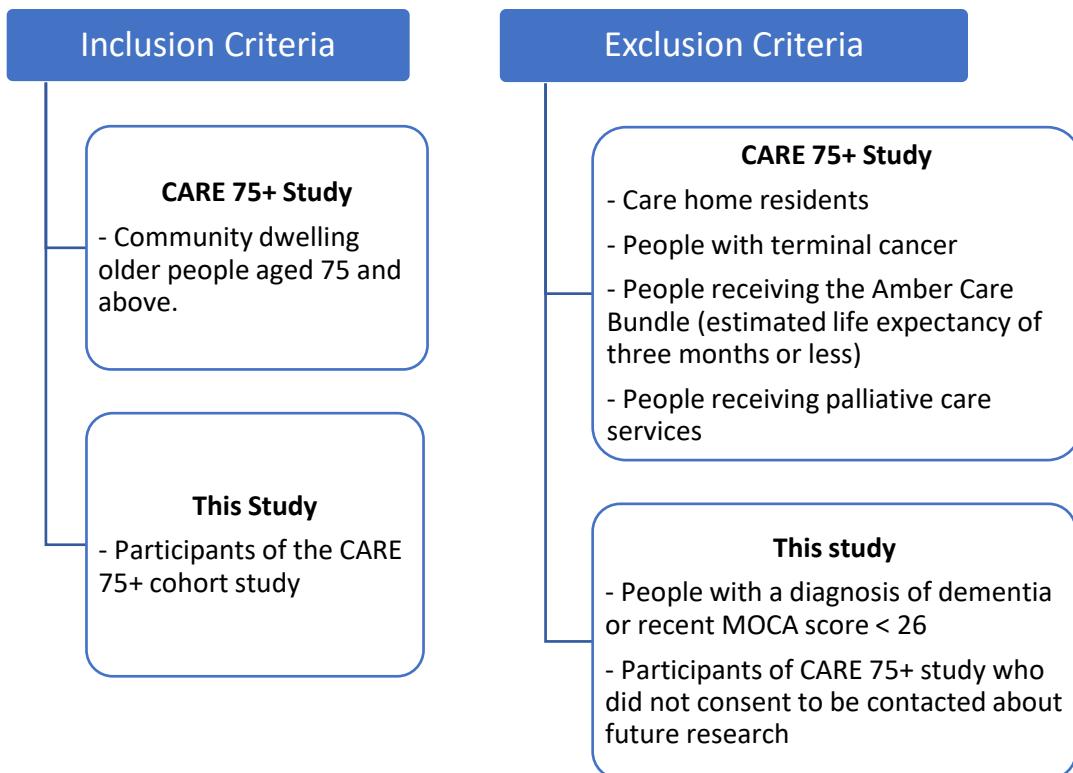
Participants were recruited from the ongoing CARE 75+ cohort study (National Institute for Health Research, 2014) using convenience sampling methods. Participants were initially recruited to the CARE 75+ cohort through GPs in Bradford and Leeds. This has been extended to various sites nationwide, however recruitment for this study focussed on CARE 75+ participants living in the Bradford and Leeds area. Only those who had consented within the CARE 75+ study to be contacted about future research projects were invited to participate in this study. As participants

were recruited from an existing cohort study, they were subject not only to the inclusion/exclusion criteria of this study, but also to the criteria of the initial CARE 75+ cohort study. The inclusion/exclusion criteria of both studies are shown in Figure 43.

All participants were over the age of 75, with varying levels of frailty between fit and frail, defined by the Fried measure of frailty (Fried, Tangen et al., 2001). All participants had a recent Montreal Cognitive Assessment (MoCA) (Nasreddine, Phillips et al., 2005) score of at least 26, the widely recognised cut-off for normal mental capacity (Davis, Creavin et al., 2015), indicating sufficient mental capacity to consent and comprehend the tasks required in the interview. Fried and MoCA scores are both assessed in the CARE 75+ study. Therefore, the most recent scores of potential participants (obtained within the last year) from the CARE 75+ assessments were taken and used in screening and sampling patients. The researcher did not administer either of these assessments in this study.

In the literature, cognitive content validation studies such as this vary in terms of sample size. This is often dependent on the type of interview style chosen. A sample size of 7-10 one-to-one interviews has been suggested to be sufficient to confirm respondents' understanding of items and concepts (Willis, 2005). An investigation into the content validity of the Dutch versions of the ASCOT, EQ-5D-3L and ICECAP-O in older people using cognitive one-to-one interviews reported reaching saturation and ceasing recruitment after 10 participants (van Leeuwen, Bosmans et al., 2015b). However, it has been argued in the literature that the required sample size is dependent on the complexity of the PROM and the diversity of the population of interest (Patrick, Burke et al., 2011b). It is therefore difficult to predict when data saturation will be reached. In line with the studies above, it was anticipated that each measure being discussed by 10-15 participants would lead to data saturation. This is also in line with the COSMIN checklist, used to assess study quality in the literature review (Mokkink, Terwee et al., 2010), which judged sample sizes of qualitative investigations of content validity to be of excellent quality as long as at least 10 participants were interviewed. Since each interview will only discuss two of the four measures this would result in an overall sample size of 20-30 participants. However, this sample size was flexible, with data collection planned to continue until the researcher felt data saturation had been reached, with no new themes being developed through further interviews. Convenience sampling within the CARE 75+ cohort was used to recruit older adults for interview or focus groups.

Figure 43 – Inclusion and Exclusion criteria for both the CARE 75+ cohort study and this PhD study



This qualitative section of the data collection was approved by the Health Research Authority and South West Frenchay NHS Research Ethics Committee in December 2017. This project was sponsored by the University of Sheffield. Documents relating to the ethics and sponsorship approval can be found in Appendices 23 and 24. Once ethical approval was received a request was sent to the CARE 75+ research team for the contact details of eligible CARE 75+ cohort members from the Bradford and Leeds area, who had consented to be contacted about future research projects. A list of potential participants, their contact details and their frailty score as defined by the Fried scale (Fried, Tangen et al., 2001) was received in February 2018. Convenience sampling was used to recruit older adults from this list.

Invitation letters were sent out in February 2018. These were accompanied by the participant information sheet and a copy of the consent form which the participant would be required to sign if they did decide to take part (shown in Appendices 25 and 26). Potential participants were instructed that if they were interested in participating in this project, they could either call or email the researcher or complete and return the response card in the included addressed and stamped envelope and the researcher would contact them to discuss the project further. If no response was received within 16 days another letter containing the same information and

documents was sent out, with the same instructions. Once the researcher had received a response from individuals indicating that they were interested in participating, the researcher discussed the project further with the participant over the phone. If they wanted to participate, they were given the choice between interview or focus group and arrangements were made. If no response was received after the second letter, the individual was not contacted again, and their details were confidentially destroyed.

The presence of cognitive impairment is an important risk in an elderly population. Cognitive impairment and diagnosis of dementia were exclusion criteria in this study. All participants were recruited through the ongoing CARE 75+ study. As highlighted earlier, the ongoing CARE 75+ study interviews participants at 6 monthly - yearly intervals. As part of this interview, they administer the MoCA (Nasreddine, Phillips et al., 2005). Those with a most recent MoCA score below 26 were considered to have below normal cognitive ability and be ineligible for recruitment. Therefore, it was unlikely that a lack of capacity to consent would be an issue in those potential participants who were invited to participate in the study. As an extra precaution, during recruitment contact and prior to starting the interview, the researcher engaged the potential participant in discussions to determine whether they had capacity to give informed consent. In line with the Mental Capacity Act 2005 Code of Practice (Department for Constitutional Affairs, 2007), the researcher assessed whether the potential participant: understood the relevant information about the study; could retain the information long enough to make an informed decision; was able to use and weigh up the pros and cons to come to an informed decision and communicate that decision. If the researcher felt the participant was capable of these things, they were judged to have capacity to consent. Where it was judged that they did not have capacity ($n=1$) the researcher did not conduct the interview. For those who were unable to read or sign the consent form due to impairments, but who had capacity to consent, audio recorded verbal consent was taken prior to the commencement of the interview.

5.3.4 Analysis method

The researcher transcribed verbatim all audio recordings within five days of the interview. Data analysis was carried out alongside data collection, so that the data saturation point could be monitored. Transcripts were entered into NVIVO version 11 (QSR International Pty Ltd, 2013) for data management and analysis. The interviews

were initially analysed using thematic analysis. Thematic analysis is “a method for identifying, analysing, and reporting patterns (themes) within data” (Braun and Clarke, 2006, p6). The organisation of themes allows for a rich description of the dataset and aids and develops the interpretation of the research topic (Braun and Clarke, 2006). Transcribing the interviews was the first stage of analysis and re-familiarisation of the data. Following the principles of Braun and Clarke (Braun and Clarke, 2006), the researcher read the transcripts in full, noting any initial themes which appeared in the data. Next, a more in-depth analysis was conducted, coding sentences and phrases which either further enforced the initial open coding or generated new codes to explore. Codes and themes were identified in relation to the way participants conceptualised QoL and wellbeing and factors that were important to these concepts, as well as response issues identified in the way participants responded to items, their opinions on the relevance and acceptability of items and the format and comprehensiveness of measures. Any quotes which appeared highly significant or exemplified a key concept were highlighted for future reference. The codes were reviewed, compiled into themes and then defined, so that every theme was both distinctive but also relevant to the research.

When analysing the content validity issues identified for each item, a report was created per item that listed the verbatim transcription of each respondents’ comments on that item, as suggested by Knafl et. al (Knafl, Deatrick et al., 2007). An initial framework of response issues was adopted from the content validation of the Dutch translations of the EQ-5D-3L, ICECAP-O and ASCOT in older people (van Leeuwen, Bosmans et al., 2015b). This framework, based on the Tourangeau model of survey response (Tourangeau, 1984), was adapted during data collection. Additional codes related to the format of the measures, the relevance and acceptability of items and comprehensiveness of measures were incorporated into the framework as additional categories. The resulting framework can be seen in Table 36.

Table 36 – Coding Framework

Response issue	Definition
Practical Completion	
Length of measure	Respondent feels there are too many or too few questions
Layout of measure	Layout confusing - certain questions easily missed or not easily understood
Comprehension / Understanding	
Odd wording	Respondent finds terms/phrases unusual or odd
Difficult wording	Respondent unfamiliar with terms/phrases or struggles with a complicated structure
Recall	
Wrong time period	Participant's answer does not align with the stated time period in the measure/question
Interpretation	
Difficult interpretation of item	Respondent expresses that they do not know or understand the meaning of item
Wrong interpretation of item	Respondent interprets the item differently than what was intended by the developers
Narrow interpretation of item	Respondent focusses on one aspect of the construct or is unsure about the focus of the item
Response Option Selection	
Different answers for different aspects of item	Respondent feels that different response options apply to different aspects of the construct
Response options partly applicable	Respondent indicates that one part of the response option fits their situation and one part does not
Irrelevant response options	Respondent doesn't want to answer any of the given options
Missing intermediate response options	Respondent feels there is a gap between two consecutive options
Similar response options	Respondent feels two options are similar
Disagreement with order of options	Respondent does not agree with order of options
Inconsistent response	Response option chosen did not match what the participant said or their situation
Acceptability	
Item inappropriate/unacceptable	Respondent feels that a question is inappropriate and should not be asked
Relevance/ Comprehensiveness	
Similar items	Respondent could not see the difference between items or thought they were excessively similar
Item irrelevant	Item not relevant to the QoL of the respondent
Important aspects of QoL missing	Respondent feels that the measure misses important aspects of QoL

In order to identify where the interpretation of respondents differed from the intended meaning of measure developers, the researcher searched for concept guides, which provided details regarding developers' definitions of terms and items included in the measures. A full concept guide has been previously published for the EQ-5D-3L (Brooks, Rabin et al., 2003), shown in Appendix 27. A partial concept guide was identified from the WEMWBS website in a document which outlined and resolved some common issues identified in translating the WEMWBS (WEMWBS Research Team), shown in Appendix 28. This document did not cover all included concepts and important terms from the WEMWBS however, it did provide some helpful clarification around intended meaning of some of the less obvious terms. For the SF-12 and ONS-4, no concept guides could be found which clarified the exact intended interpretation of items. An SF-36 concept guide was also searched for in the hope that it would provide intended interpretations for the SF-12 items, but none could be found. For these two measures, as well as the concepts of the WEMWBS not covered by the document identified, the researcher examined previous published development and validation literature about the measures in order to deduce the intended meaning of developers.

5.3.5 Reflexivity

In all qualitative research there is the need to reflect on the role of the researcher in the research. This is because the researcher's background and knowledge base will influence the interactions that they have with their participants. For instance, it was observed that participants would sometimes seek approval for the responses they gave. For example, Mrs Eight asked '*Were my answers ok?*'. This shows a tailoring of responses which they believed would fit the aims of the researcher. This is opposed to more spontaneous and less constructed answer. However, these responses often reduced during the interview when participants began to feel less anxious about their responses.

Despite it being made clear in the introduction to the interview that these were not surveys that had been developed by the researcher and that all opinions about the questionnaires, both positive and negative were welcome, some participants would ask part way through the interview if these were questions the interviewer had designed. It was clear that they were reluctant to criticise something that the individual in front of them may have made. Once reassured again that these were existing

surveys, that had not been made by the researcher or anyone affiliated with the researcher and that the aim of the study was to understand what they truly thought of the questionnaires, participants seemed happier to voice negative opinions.

One thing that distinguished the researcher most from participants was age. This was often brought up in interviews, when participants would give examples of how things differed between their age and the researcher's age. However, it was not felt that this necessarily had a negative effect on the research. It often meant that participants explained themselves and the experience of ageing more fully than they would have to someone closer to their age as they did not assume that the researcher had prior knowledge or understanding. They also provided useful examples of how what was important or relevant to them differed between when they were the researcher's age and their age which may not have occurred to them in a conversation with someone of a more similar age and these examples were often useful and reinforced important findings.

5.4 Results

5.4.1 Recruitment and respondent characteristics

A total of 122 potential participants were contacted by letter. Forty responses were received, of which 22 agreed to take part and 18 declined either due to not being interested in taking part ($n=7$), ill health ($n=5$) or ill health of partner ($n=2$), too busy to take part ($n=1$), no longer at the registered address ($n=1$) or a family member reporting that the person being contacted had passed away ($n=2$). The vast majority of responses were through the response cards, with only two telephone responses and three email responses being received. One participant who sent a response card expressing interest in participating could not be contacted. One participant was consented, but during the interview it became clear that their mental capacity had declined, and the researcher could not be sure that they had sufficient mental capacity to consent and therefore the interview was stopped, and data provided up until that point was excluded from the analysis.

Recruitment and data collection were stopped as saturation was reached after the 20th interview. Twenty participants were interviewed in full and included in analysis. All interviews were one-to-one, with the exception of a married couple who were both members of the CARE 75+ study and wanted to be interviewed together (Mrs Four and Mr Five) and one participant who requested her daughter be present at the interview (Mrs Thirteen). Two participants had very poor vision (Mrs Eight and Mrs Ten) and requested that the measures be read out loud to them. All participants completed both measures asked of them in a single interview.

The characteristics of the sample as a whole are described below in Table 37. Thirteen of the participants (65%) were female, fifteen (75%) lived alone while the remaining five (25%) lived either with their spouse or child's family and four (20%) lived in an assisted living facility with a warden. The average age (range) of the participants was 83.95 (77-94). Fifteen (75%) reported at least one long-term condition (LTC). According to the Fried scale, at their last CARE 75+ assessment nine (45%) of the participants had been classified as pre-frail and eleven (55%) had been classified as frail. The sample for this study were more likely to be female, over the age of 85 and living alone than the average in the general population aged 75+. Our study sample were also healthier than the general population aged 75+, according to the EQ-5D-5L, but also more likely to be classed as frail. The characteristics of each individual participant, as well as the combination and ordering of the measures they discussed in their interview are shown in Table 38.

Table 37 – Qualitative participant characteristics

Characteristic	Number	Percentage of sample	General Population Aged 75+ Comparison (reference)
Number of participants	20		
Female	13	65%	58% (a)
Average Age (range)	83.95 (77-94)		81.95** (a)
Aged 75-84	11	55%	71% (a)
Aged 85+	9	45%	29% (a)
Average Fried Score	2.85		
Fried Fit (score 0)	0		
Fried Pre-frail (score 1- 2)	9	45%	
Fried Frail (score 3-5)	11	55%	Approximately 30% (b)
Reported any LTCs	15	75%	
Live Alone	15	75%	38.5% (a, c)
Live in assisted living facility	4	20%	
Ethnicity - white	20	100%	
Measure Scores	Mean	SD	
EQ-5D-5L Utility	0.80	0.17	0.734 (EQ-5D-3L) (d)
EQ-5D-5L VAS	82	7.89	73.8 (EQ-5D-3L) (d)
WEMWBS	51.2	9.52	50.96 (e)
ONS-4*	31.8	5.51	

a (Office for National Statistics, 2018). b (Gale, Cooper et al., 2015). c (Office for National Statistics, 2017a). d (Janssen and Szende, 2014). e (Davidson, Sewel et al., 2009)

* ONS-4 score generated by reversing the scores on the anxiety question and then summing across items so that higher scores indicate higher levels of wellbeing.

** Average age (year 2016) may be slightly underestimated as life tables present each age up to 105+ so it was assumed all these people were 105. However, there are only 874 people aged 105+ out of a population of individuals aged 75+ of 5,325,503 (0.02%) so the impact will be small.

Table 38 - Measure Combinations and Individual Characteristics

Combinations	Who	Age	Fried	LTC	Quals	Lives alone
EQ-5D-3L - WEMWBS	Mr One	75-79	1	Y	None	Y
	Mrs Eight	85-89	3	Y	None	Y (Assisted Living)
WEMWBS - EQ-5D-3L	Mrs Seven	85-89	5	Y	None	Y
EQ-5D-3L - SF-12v2	Mr Two	90-94	4	Y	None	Y
	Mrs Twenty	75-79	4	Y	None	Y
SF-12v2 - EQ-5D-3L	Mrs Fifteen	90-94	3	N	None	Y
EQ-5D-3L - ONS-4	Mrs Eleven	80-84	4	Y	None	Y (Assisted Living)
	Mr Fourteen	75-79	2	Y	Postgrad	Y
ONS-4 - EQ-5D-3L	Mrs Four	80-84	1	Y	None	With husband
	Mr Five	75-79	2	Y	None	With Wife
WEMWBS - SF-12v2	Mrs Nine	80-84	3	Y	GCSE	Y
	Mrs Eighteen	80-84	1	N	None	Y (Assisted Living)
SF-12v2 - WEMWBS	Mrs Sixteen	90-94	2	N	None	With child's family
	Mr Twelve	75-79	3	N	None	With Wife
WEMWBS - ONS-4	Mrs Ten	90-94	2	Y	None	Y
	Mr Nineteen	85-89	4	Y	None	Y
ONS-4 - WEMWBS	Mrs Thirteen	90-94	3	Y	None	Y (Assisted Living)
SF-12v2 - ONS-4	Mrs Three	85-89	2	N	None	Y
ONS-4 - SF-12v2	Mrs Six	75-79	3	Y	None	Y
	Mr Seventeen	75-79	2	Y	None	With child's family

5.4.2 Findings

5.4.2.1 Quality of life and Wellbeing definitions

The discussion of the definition of QoL and wellbeing and what factors were needed in life to achieve a good level of these were analysed thematically, separately to the content validation of the measures themselves. It was felt that it was important to understand respondents' opinions of what QoL and wellbeing meant to them before the content validity and response issues found for the measures could be fully understood. Whilst QoL and wellbeing were difficult for participants to differentiate, these concepts were often linked to the following broad themes: health, ability to carry out usual activities, social participation and emotional functioning.

Health

Health was discussed as an important element of QoL by all participants. Aspects of health which were commonly mentioned were specific health conditions, mobility, pain, cognition, energy. For example, when asked what was important to his QoL, Mr Fourteen responded "*Well no surprise at all, the cliché good health and ability to be mobile and to think very clearly and to feel useful really. Those things seem to be very important.*" However, the way that these aspects of physical and mental functioning, as well as general health were discussed revealed that it wasn't health in itself that was viewed as important to participants. It was the impact that health had on their ability to undertake activities that they valued or enjoyed and their ability to access and participate in regular social interaction.

"If you've got your health at my age, you don't need a lot more in life because you can get out and about and do stuff. So, god help me, I hope I never get to that stage where I can't go out." Mrs Eighteen

The dialogue around the topic of health revealed important findings about the way older adults think about and assess their own health. People's view of their own health and QoL was often assessed relative to others they knew in a worse state. Their comparators were often friends or family members of a similar age to them. They would often use these people as an example of how lucky they were to be in a relatively better state, despite their own problems. This often led to them describing

problems with their own health but then going on to view their own state very positively, as they were not as badly off as others around them.

"My friends that go to these classes, they're none of them better off than me. Some's a lot worse. I mean this lady who's not well, she's (the same age as me) and to look at me and to look at her.... Terrible yeah. Yeah, she's carers going in and she must have Alzheimer's, but she's very witty and you've got to laugh. She's a lot different to my carry on and mines a lot better. You've got to think on the good side." Mrs Seven

This is an important finding in relation to the way older adults report their own health status. The consideration of health state as a relative concept (relative to other people of a similar age who are in worse health), rather than fixed value could have a substantial positive impact on the scores older adults provide for their health. This could be an example of response recalibration if, as older people age and their health and ability to function independently declines, they shift their benchmark for what constitutes good health downwards. By using others of a similar age who are worse off as the new reference they can continue to rate their lower state positively. This is an issue as if different individuals use different strategies to assess their health, or make assessments based on different references, the scores they provide are not necessarily comparable.

Participants' view of their health was also very strongly linked to the process of ageing and declining health. It was clear from the way that many respondents spoke about their health that their expectations of their health were lower now than they had been when they were younger. People spoke about "expecting" their health to have declined in their old age and they expected this to continue. An example of this was Mr Two, who said *"I expect to go down a bit. You don't expect to stay the same active as you were 10 years ago and at the moment, I've put 75%. Well in a fortnight I could drop down to 50%, for all I know. Anything can change as you're getting older... so quick."* Declining expectations also have an important impact on the scores that older adults provide. By assessing their health based on what they "expect" for their age, which is lower than what they would have expected when they were younger, their benchmark for what constitutes good health shifts downwards in older adults. This again provides an example of response recalibration. Again, this means the scores they provide for a given health state are not likely to be comparable with those

provided by younger adults, as older adults are more likely to view that state more positively.

It was also common for participants to mention uncertainty about their future health, the speed with which their health and level of functioning could change at their age and death. They often referred to the fact that their health could change and decline very quickly and without warning, such as Mr One who said "*It doesn't affect me really at the moment, but anything can change. Same as I tell doctors. At our age you can go like that (clicks fingers).*" Many respondents mentioned the prospect of them dying and when or how this might happen. Uncertainty and the fact that their level of health and functioning could change so quickly at their age were usually key parts of discussions of their deaths. Their own death was often mentioned in an accepting way, as if they expected it to come and did not fear this. The death of loved ones and friends was discussed much more emotively, with some participants becoming visibly and audibly upset. How they would die, or the state they would live in before that point, seemed more important than when they would die and often respondents would refer to states that they would not want to live in.

"The older you get, the more you come to realise that it's not forever. That you're going to go at some time or another. When you're young like you, you don't think about it. But you think about it when you get to my age." Mrs Twenty

"I worry about how long I'm going to be mobile – as mobile as I am now, because my walk is definitely deteriorating. I go to exercise classes once a week. That does improve you to a certain extent, but not brilliantly. I've got a walker out there, but I don't like it. My daughter says I don't like it because it shows I'm old – she may be right too. But no, I worry about the future in that what's it got for me. I don't want to go to a care home because I love my independence too much for that and my daughter has assured me that that won't happen unless I want it so that's alright." Mr Nineteen

While, death was usually discussed in an accepting manner, uncertainty about the state of their future health and the impact this would have on their ability to live independently was a commonly expressed concern.

Health was considered an important aspect of QoL by all participants as it impacted their ability to undertake activities that they value and enjoy and to participate in regular social contact, which were both of central importance to their view of their

QoL. The way health was discussed revealed some important insights into the way older adults view, think about and assess their QoL, which will impact the responses they provide on health-related PROMs. They commonly expressed that their expectations of their health had declined with old age. It was also common for respondents assess their health as relative to individuals of their age who were in a worse health. Both these response mechanisms provide evidence for response recalibration as respondents shifted the benchmark for what constituted good health downward, meaning they view their health more positively than would be expected.

Ability to carry out usual activities

People's ability to carry out their usual activities was clearly important to their QoL. This included both daily household activities, usually described as self-care, housework and gardening, as well as activities outside the home such as getting out and about, going to social clubs and meeting with friends and family. Respondents' ability to carry out their usual activities was dependent on health-related factors such as mobility, energy and pain. Their ability to undertake these activities was commonly discussed in the context of independence, support, adaptation and confidence.

Mobility, energy and pain were often discussed in terms of their impact on participants' ability to undertake regular tasks. People reported that they had less energy and they felt themselves slowing down. This meant that, while they could often still do "things", as they broadly described in their interviews, it took much longer, and they found it much more tiring than they would have done at a younger age, such as Mrs Fifteen who said, "*I can do most things put it that way. I'm sometimes slower doing them but who's to bother, there's only me so it doesn't matter*". Reduced mobility and pain also had an impact on respondent's ability to carry out their daily activities around the home as well as access activities outside the home.

"I get horrible pain in my back and it goes right across my back and it's like toothache and I can't walk, I can't. It cripples me. And I have to sit down. I can't even get from here to the bus stop because it's so bad." Mrs Twenty

"I mean I might hoover up and then sit down a bit or make my breakfast. I mean, I do bits in between and then sit a while. It might take me all day if I'm in, but I get there. But I like to do it myself." Mrs Eleven

As their health and ability to carry out their daily activities declined with age, there was an obvious desire to remain independent where possible and to do household activities, such as the hoovering, for themselves. For example, Mrs Eighteen expressed frustration at the suggestion of others that she should “take it easy at her age” saying *“It’s always ohhh, at your age, at your age. Well I don’t want to be sat around at my age, I want to be doing things [...] I try to do them myself, because I like doing stuff myself”*. Remaining independent, and discussion of areas where they were still independent, was clearly a source of personal pride for participants.

Finding ways to continue performing their daily activities independently in the face of declining physical functioning was often linked to adaptation, as people sought out ways to make things easier for themselves. For example, Mrs Fifteen reported, *“Every time I slip or anything its ohhh, we’ve got to find some way of getting round it.”* This ranged from making jobs less physical, such as sitting down when ironing, to making adaptations to the home, such as stair rails.

Only once more physical tasks became too difficult would respondents seek support and help. Their support network commonly consisted of friends and family, but as their support needs increased for frailer participants, more formal support was necessary. The theme of support and help was very strongly linked to the desire to remain independent where they could. Mrs Eleven expressed this when she said *“I don’t want folk mollycoddling me. I want to do it myself. I know I struggle, I mean, but I get there in the end. At times it might be too much, I don’t know.”* Participants often expressed frustration, which arose from the internal conflict between the desire to remain independent but the knowledge that they needed support. Participants were keen to focus on maintained abilities rather than their limitations and always emphasised their independence over any need for support. Often, when a participant mentioned an area where they needed help, they were often quick to mention the things they still did for themselves.

It was very important to people that they were able to get “out and about” and engage in activities outside the home. This most often involved visiting friends and family or attending social clubs. These activities were an important source of enjoyment and social contact. Particularly for those participants who lived alone, getting out and about was essential to be able to access regular social contact. Respondents would often describe feeling down if they were stuck in the house for days in a row.

"I need to be able to get out. [...] I love getting out. I always... I couldn't bear stopping in unless the weather was bad and then I wouldn't go out. But I go out most days, even if it was only just to go and look round the shops or somewhere. But I used to love going to Yeadon and getting on the bus to Harrogate. [...] I just like getting out and I have got uhh friends down below in the bottom part of (the village) and we meet up and so that is nice as well."

Mrs Sixteen

"So as long as I can get about. It's when I'm stuck in and I can't get out, it gets me down. So, I don't know how I'll get on when I get older. But anyway, I'm doing alright. I try to keep my spirits up, you know what I mean." Mrs Twenty

Being able to get out and about was also linked to keeping busy, feeling involved and having social activities to look forward to, all of which people felt were important aspects of a good QoL. People described the importance of keeping busy and feeling involved with family, friends and the community and how this impacted the extent to which they could feel useful and wanted. The concept of making the most of life was also frequently mentioned by respondents. As they saw their health and ability to do their usual activities decline, they felt it was important to make the most of their current abilities, as they did not know when they would decline further.

"I try to make the most of every day. Uhh I say you don't know what's going to happen tomorrow, so make the most of it when you have it, you know. And I try to do that as much as I possibly can. I enjoy as I say, I enjoy reading and I enjoy watching programs on the telly as everybody does generally. But I'm thankful to be here really and truly at my age." Mrs Sixteen

"My (grandchild) gets married next year and I said, I might not be here, and they said you better be grandma (Laughs). You know, and I think when you've something to look forward to, its better isn't it. You're not thinking ohhh it's just going to go on and on are you. But yeah, I think, oh yeah there's that to look forward to and they involve me a bit anyway which is good. I like something to look forward to." Mrs Eleven

People also felt that having social activities and family events to look forward to, gave them a reason to carry on being active, as mentioned by Mrs Eleven above. Without these aspects of QoL people feared they would become insular, get left behind and lose a sense of worth in life.

Respondents' ability to get out and about was dependent on their mobility, energy and their local environment. Participants felt that their energy levels varied a lot day to day, and this had an impact on when and how often they felt they could engage in activities outside the home. Their mobility affected how far they could get and the level of activity they could undertake. As mobility and energy declined, the facilities and accessibility of the local area became increasingly important. This is an example of response reprioritisation, with the home and local environment gaining in importance as physical functioning declined. For example, Mrs Six, who had recently bought a mobility scooter, as she could no longer walk far, said "*Me and my friend we go out for meals you see at different pubs that do meals and then we go both on the scooters together, so we can get a bit further. But I use buses a lot – I'm always on bus.*" Many respondents, including Mrs Six, relied heavily on buses to access activities outside of the home. For those who drove, driving was seen as important to enable them to easily access the local area and activities. This was clear from comments such as "*I've got mobility with my car [...] if I didn't have my car I wouldn't go out and get as far.*" from Mrs Eleven. Giving up their driving licences or cars was often mentioned as a loss of independence but often driving was an area where they lost confidence in their own ability.

Significantly, as people's health declined, they mentioned losing confidence in themselves and their ability to undertake their usual activities.

"I bought my car up here and I kept it for a year... I've been here 3 years. But I was losing confidence in myself, especially with these youngsters coming up. So, I sold it and my son said it was best." Mr Nineteen

"You know, and the thing is I've found. It's almost now a year, because it was April when my hip gave way. My eldest daughter, because she's not working, got the car and was taking me to the hospital for appointments and that kind of thing. And I find that its then difficult, because I've been going out that way (in the car), and especially when you've gone through the hip business. There's no problem with the hip, no pain whatsoever but you lose a little bit of confidence." Mrs Three

Having control over their daily live was clearly very important to respondent's QoL. Being able to do what they wanted, when they wanted, was central to many respondents' concept of QoL or satisfaction with life. For example, when asked what

she thought about when deciding how satisfied she was with her life, Mrs Ten stated “*being able to get out and about, being able to do what I like yeah in my own house*”. The concept of control over daily life was very closely linked to remaining independent.

Peoples’ ability to carry out their usual activities both inside and outside the home were of central importance to their QoL. Their ability to carry out their daily activities within the home was closely related to their sense of independence and pride. Areas where they were still able to be independent were always proudly emphasised over areas where they relied on support, which was often a source of frustration. Their ability to engage in activities outside the home independently was closely related to their sense of control over their life and their ability to engage in social contact. As their mobility and energy levels declined and they experienced pain, the accessibility of their home and the local area became increasingly important for them to continue independently.

Social Participation

Regular social contact of various types was obviously central to peoples QoL, with all respondents mentioning some form of social contact when asked what they needed in life to feel they had a good QoL. Family, partner/spousal relationships and friends were all important sources of social contact. Those who did not have many friends or family members close by described speaking to people in the community whilst out and about or chatting with carers or people who came to the house. While some participants said they were fine alone and did not need constant social contact, everyone mentioned some form of regular contact with other people. Loneliness was often mentioned as a big problem in older people.

“Ohhh, I feel downhearted sometimes. When you’re on your own. I’m not as bad now with (neighbour) coming in but at one time, when she wasn’t coming in... I used to go with her shopping, but she didn’t call in. Now she calls in every day. It makes a big difference. A very big difference. People don’t realise how much difference somebody calling in makes to a person on their own. It makes a hell of a difference. It does. It’s the most important thing. Loneliness.”

Mr Two

"I think I've just about got everything ummm my family... I have a very outstanding family. And I've got good neighbours. Uhh they take me about, they don't leave me on my own very much. Usually every day I have somebody popping in or out, you know, or I'm going out." Mrs Fifteen

The importance of feeling involved with their family was clear. Some mentioned feeling that they were a burden on their family. The emotional impact of this was seen and heard during the interviews as they looked visibly upset when they responded saying they felt saddened and uninvolved.

"But I wish that I had more family life. My eldest daughter is [living far away] and my other daughter, over there (points in direction of daughter's home), is very good with me, but I don't feel as though I'm part of the family. [...] I just feel a bit out of it sometimes but there's lots worse. There's lots that haven't anybody have they, anybody at all. It's just me being bitchy I think. [...] But I do feel that... that when she comes, and we go out I feel that she's thinking it's a duty, which it is" Mrs Seven

Feeling that they were a burden on their family was a clear concern of many participants. It was a feeling they wanted to avoid, and they often stated that they tried to remain as independent as possible so that they did not have to feel this way.

Emotional functioning

The emotional impact of ageing and the strategies used for coping with this were also often referred to during discussion about their QoL and wellbeing. Participants often expressed feelings of frustration in relation to their declining ability to undertake activities independently and concern about the uncertainty of their future. Commonly mentioned strategies for coping with the emotional impact of ageing surrounded themes of stoicism and the importance of having a positive outlook.

While some participants were very accepting of the ageing process, several expressed frustrations at their declining functional abilities. For example, Mr Nineteen responded *"I don't say very much about it, but I don't like being old. I don't want to be young though. I just want to be normal, you know, have my facilities like. Physically I'd like to be more active. I mean you consider that 40 years ago I was taking boys up Snowdon and I can hardly get up the curb now (laughs)."* Frustration with ageing was

most commonly expressed in terms of no longer being able to do activities which participants had enjoyed and having to rely on others for support rather than being able to achieve things independently.

There was a strong theme of stoicism running through the interviews, with the idea of enduring issues and hardship without complaint or showing feelings. At some point in the interview most participants expressed, in some form, the importance of not dwelling on or worrying about things that could not be controlled, as this was not good for them and therefore it was better to just carry on with life. For example, Mrs Nine made several comments, when discussing negative events such as "*I'm not into emotion noo it doesn't worry me at all really. It... you have to get on with it, I'm sorry. I've always tried to be practical*" and Mrs Sixteen said "*I'm not one that dwells on things. [...] I try to look on the bright side as much as I can*". This stoic attitude was strongly linked to the idea of positive thinking and looking on the bright side as a way to carry on.

People often expressed that not having to worry was an important part of QoL. Common potential sources of worry were financial problems, family and friends not doing well, dealing with problems and their future living situation and ability to live independently. Money was often mentioned in relation to QoL.

"We could do with a bigger pension. You know when you retire you think ohhh yeah, we'll be alright. Its uhh... It could just do with topping up a bit more. [...] not a lot of money. Just enough to be able to do nice things together, to get out and about." Mr Five

Respondents, such as Mr Five, expressed that they felt they needed enough money to be able to do all the activities they wanted to do, without having to worry about money and their future financial security. The happiness and health of family and friends was also a common concern. Having to deal with problems, for example with the home, was a worry for some who felt that they had lost confidence in their ability to fix things or organise solutions. Finally, the future was sometimes a source of worry for participants in terms of their future need for support and living situation.

"I worry about the future in that what's it got for me. I don't want to go to a care home because I love my independence too much for that and my daughter has assured me that that won't happen unless I want it so that's alright." Mr Nineteen

Security about the future, both financial and regarding the need for support and whether the type of support that the individual would want would be available, clearly concerned some respondents. This often stemmed from a desire to remain independent. Support networks, such as family members were often mentioned in relation to such concerns. However, mostly respondents preferred to think positively and stoically carry on with life as they did not feel that worrying about issues that they could not control helped them.

Wellbeing Definition

Most participants did not feel there was a difference in the meaning of QoL and wellbeing and thought they captured very similar, if not the same concepts. Some participants considered wellbeing to be more closely linked to health. In at least one case this was due to contact with health services with the term wellbeing in the name of the service. This led to several participants feeling this term was more official, such as Mr Nineteen who said "*Wellbeing sounds more official, more institutional somehow I don't know why, but it does to me. It sounds like one of these words that have come in since the war.*" The definitions of wellbeing given by several participants suggested that to them wellbeing literally meant being well, such as Mrs Seven who said, "*Wellbeing's your health isn't it?*". One participant, Mr Fourteen, felt that wellbeing was a more subjective concept than QoL, as QoL was more objectively seen and measurable by what you could and could not do. Mr Fourteen was still working in connection with health services and this may be why he was more confident in giving a definition of wellbeing as different to QoL.

"I think it's about meaning the same sort of thing (as quality of life), that you're happy within yourself... things are going alright. It could be different, and you could be needing help and things like that. But no, I'm quite happy." Mrs Fifteen

"Well, wellbeing is very much a self-perception whereas quality of life can be judged by an external. So, for example, a doctor might say uhh you know you're mobile and, and you know, you still get out so that is quality of life. But wellbeing, perhaps, is things like... as far as I can tell what many older people in particular encounter is a feeling of being somewhat marginalized." Mr Fourteen

5.4.2.2 Content validity

In the following section the content validity of each item from each measure is assessed. The response issues identified for each item during respondents' think-aloud completion and subsequent verbal probing are presented and the threat these response issues present to the content validity of the measure are examined.

EQ-5D-5L

Response issues identified during respondents' think-aloud completion of the EQ-5D-5L and subsequent verbal probing are presented in Table 39.

Measure as a whole

People found the layout, style of the questions and response options easy to understand and answer and nobody had a problem with the length of the questionnaire. The EQ-5D questions, particularly the first three, focus on people's view of their functional ability and it was clear that people found this type of question easier to answer than more subjective questions. Respondents also felt it was relevant to a wide range of respondents at different levels of health.

"You can pick out what suits you. Yeah it was very good that one... It was specific in what it was meaning." Mrs Fifteen

"Yeah because I think that, as I say, you don't really have to think about it too much. You see these, well ability again, you don't have to think about it because yes, I can walk about, that's it. Whatever the question, it's perfectly clear which box to tick." Mr One

"They were all, you know, perfectly normal questions to ask anybody and whether they're, you know, confined to a wheelchair or active, it covers everybody does this doesn't it... Because you've got it here that if there's someone like my wife (confined to a wheelchair), you can say they've got problems. You know, its scaled to suit everybody and then you can also get such as me in, who says yes to everything, no problems to everything" Mr Five

Table 39 – EQ-5D-5L Response Issues

Each / represents a participant who experienced the corresponding response issue

	EQ-5D-5L						
	Measure whole	Mobility	Self-care	Usual Activities	Pain/discomfort	Anxiety/depression	VAS
Response issue							
Practical Completion							
Length of measure							
Layout of measure							
Understanding							
odd wording							
difficult wording							
Recall							
Wrong time period		/					
Interpretation							
difficult interpretation							
wrong interpretation			/				
narrow interpretation						///	
Response Option Selection							
Format difficult							
Different answers for different aspects of item		//					
response options partly applicable							
irrelevant response options							
missing intermediate response options							
similar response options							
disagreement with order of options							
Inconsistent response	///	/	/				
Acceptability							
Item inappropriate							
Relevance/ Comprehensiveness							
Similar Question							
Item irrelevant						/	
Important aspects of QoL missing	///						

"Yeah because I think that, as I say, you don't really have to think about it too much. You see these, well ability again, you don't have to think about it because yes, I can walk about, that's it. Whatever the question, it's perfectly clear which box to tick." Mr One

"They were all, you know, perfectly normal questions to ask anybody and whether they're, you know, confined to a wheelchair or active, it covers everybody does this doesn't it... Because you've got it here that if there's someone like my wife (confined to a wheelchair), you can say they've got problems. You know, its scaled to suit everybody and then you can also get such as me in, who says yes to everything, no problems to everything" Mr Five

From these quotes it was clear that participants liked the wording of the EQ-5D items and response options because they were clear, unambiguous, easy to understand and the response options were distinct, easy to choose between and suitable to cover a broad range of respondents. The focus on functional ability also made the questions easier to answer for respondents as they did not have to think too hard about subjective concepts that they did not necessarily think about day to day and therefore felt confident selecting an appropriate response.

Item 1 Mobility

Everyone thought mobility was an important aspect of their QoL as it effected how well they could get out and about and carry out their usual activities, demonstrated by Mr One who responded *"Yes, well it is (important), I suppose. If you can't get about, that's it isn't it (laughs)." Nobody felt it was inappropriate to ask.*

"It is (important) yes, it is. Definitely, if you can't move about its shocking. I've put moderate, depending. Sometimes I'm better than others." Mrs Seven

There were several response option selection issues. Some participants mentioned that their mobility varied over time and depending on the situation, which sometimes led to issues when trying to pick between several potentially relevant response options. For example, Mrs Fifteen asked for clarification by saying, *"does this walking about mean in the home or outside?"* Some respondents, such as Mrs Eight below, found it easier to get around inside the house, where they were more familiar with their environment and often had adaptations to assist them, than outside where it was

more difficult. Some participants also struggled to stick to the timeframe of “today” in relation to this question. They felt their mobility varied over time, dependent on flare ups of related conditions and they therefore ignored the today statement and chose a statement which they felt reflected the more general recent state.

“Well I’m alright in here because there’s rails. I’m alright with my stick. If I’m not carrying anything, I’ll take my stick. There’s hand rails all the way down there (in the corridor) but when I go out, I just freeze. I couldn’t go out (of the building) that way on my own, but I do go down that way because there’s fence and a rail. I’d say moderate. I’m alright indoors, it’s just if I forget my stick, I open the door and freeze I’ve got to come back for it.” Mrs Eight

“Uhhh well I could tick 2 really – moderate problems walking about, but sometimes when my back’s bad, severe problems. Can I tick two? No uhhh well, at the moment, I haven’t severe problems, so I’ll put moderate problems. I wobble a lot, but I don’t actually fall. But I bump into stuff a lot.” Mrs Seven

There were also issues of inconsistent responding to this question. For example, Mrs Four who was confined to a wheelchair, selected “severe problems in walking” about rather than “unable” and Mr Two chose “slight problems in walking” about despite having said *“I used to be able to go down (walk to his allotment – about a 5-10-minute walk)– I don’t go down now because my legs are buggered. I’ve got an electric go kart yeah. I can walk but not far.”* It was not clear whether this was due to social pressure in answering or whether response recalibration meant that they had adapted their expectations based on experience of declining health and physical functioning.

Item 2 Self-care

Everyone thought self-care was an important aspect of their QoL. Nobody felt it was inappropriate to ask. It was often linked to independence. Again, there were issues with inconsistent answering. For example, Mrs Eight who is blind responded *“No problems. All I can’t do is fasten buttons.”* yet she had described several stories about adaptations she had had to make to be able to get dressed and times when she had struggled. Stories about adaptations were common in relation to this question. Several respondents compared themselves others they considered to be in a “much worse” state, who were unable to care for themselves, when answering and stated that they were very fortunate to be in a better state.

"Well yes, yes it must be a right bind if you need help to shower in the morning or something like that. In fact, I have a very good friend that uhhh, he can't do anything for himself. He needs a wheelchair to get around the house and he needs help dressing and bathing and things and when I've seen I think I'm very lucky." Mr One

"I have no problems washing or dressing, no. Sometimes it depends, you learn to do it your way, you know what I mean. Like when you're putting your pants on. I can't lift both legs... I can lift this leg fine, but I've got to hold on when I lift this leg. So, I should say moderate problems. I can shower myself as long as I can take my time... See it all depends on your surroundings, what you do. You know like I've got a walk-in shower, I've no steps and I can get about fine. It's like from here the bedrooms next door, bathroom in there, kitchen in there so it's all local all around and I can do it ok." Mrs Twenty

Respondents, such as Mrs Twenty, noted the importance of their home environment in enabling them to remain independent in self-care activities. These quotes highlight the participants' need to adapt both their ways of completing self-care activities and their home environment in order to remain independent in these activities. This was clearly important to respondents, highlighted by the emphasis brought on how lucky they were compared to others they knew who were unable to care for themselves independently. This desire to remain independent may have contributed to feeling social pressure to respond positively in relation to this question. Or experience of ageing and gradual decline in health and functioning may have led to response recalibration.

Item 3 Usual Activities

All respondents felt that being able to do their usual activities was important to their QoL and that the question was acceptable to ask. There was a common issue with narrow interpretation of this item. The EQ-5D concept guide states that the usual activities domain is intended to encompass work (paid and unpaid), study, housework, leisure and social activities (Brooks, Rabin et al., 2003). Interestingly, gender appeared significant in participant's interpretation of "usual activities". Female respondents were more likely to correlate it with being asked only about household jobs such as cleaning, cooking and gardening, whereas some male participants interpreted this question as asking about activities outside the home, even though

many respondents of both genders had mentioned being involved in such activities in other parts of the interview.

"Usual activities... do you mean cleaning and that? Well I do most of my own. Occasionally my daughter will come and say I'll hoover for you, but I do it myself mostly. But I can manage." Mrs Seven

"Well everything that I normally do. If it's a nice day I go for a walk, I go out on my bikes." Mr One

Again, there were issues with inconsistent responding. For example, Mrs Eleven reported having no problems with her usual activities but then described adaptations she had had to make to do her usual activities such as having to sit to iron and take lots of breaks. Again, this could be due to response recalibration, as the benchmark for what constitutes "no problem" in conducting usual activities shifts downward as ability to function declines with age. Many participants talked about not being able to do as much anymore as they were much slower and had had to adapt the way they did their household chores. They also often mentioned that there were things they could no longer do and required help with.

"I can do everything if I can take my time; put my own washing in, I can peg it out on the line. I can iron as long as I can sit down. If I can sit and iron on my iron board, I'm alright. Sometimes the kids will say I'll iron for you mother. I've a bit of problem making the bed, but it's my own fault because I should get a smaller bed. I've got a king-sized bed. I'm lost in it. I've had it for years. So, it takes me about an hour and a half, but I can manage. And then Jane, the cleaner, does my vacuuming and stuff for me." Mrs Twenty

Being able to manage their usual activities, particularly within the home was very strongly linked to independence with participants such as Mrs Seven making comments such as "Yes, awful if you think you can't do it" and Mrs Eleven saying "*I don't want folk mollycoddling me. I want to do it myself. I know I struggle, I mean, but I get there in the end.*" This quote provides a good example of the pride in which people took in their ability to achieve tasks independently. Despite the time and difficulty of doing so, it was clearly important to participants to continue conducting their usual activities independently where possible. This was commonly enabled through adaptation of the activity, such as sitting whilst ironing. Respondents were keen to focus on the activities they could still do rather than those they couldn't. This

social desire to be seen as independent may have contributed to inconsistent responding in relation to this question

Item 4 Pain / Discomfort

Everyone thought pain was an important aspect of their QoL. Pain was often linked to independence, mobility and being able to carry out their usual activities. Nobody felt it was inappropriate to ask.

"If you've a lot of pain you're miserable. You can't laugh that off at all." Mrs Seven

"I get horrible pain in my back and it goes right across my back and it's like toothache and I can't walk, I can't. It cripples me. And I have to sit down. I can't even get from here to the bus stop because it's so bad. But once I have the cortisone, it's alright and it'll be better still when I get my hip done tomorrow. But this is what keeps us going really. And I love it when I can walk about and have nought to worry about and I go down to the centre and I do exercises. I do exercises on a morning, not for long, and exercises on a night, before me tea. Or if not, I'll do them later at night and I feel as though I'm doing the best I can to keep mobile." Mrs Twenty

As these two participants highlight, pain was something which had a substantial impact on QoL. It impacted their mobility and ability to be able to carry out their usual activities, both inside and outside the home and participate in social contact. Pain also negatively impacted their mood. It was something that they were keen to avoid and treatments and exercise regimes to maintain movement and minimise pain were often mentioned. Patients were keen to show that they were doing as much as they could to maintain their current levels of mobility and functioning.

This was another question where issues with narrow interpretation were common. Seven out of ten participants mentioned only one of the two constructs mentioned in the question, of which six of them mentioned only pain.

Item 5 Anxiety / Depression

Most people felt it was an acceptable question to ask, but said it was not something that concerned them. People only spoke about feeling depressed at the time of the death of a loved one but emphasised that it was something they got over and hadn't had a problem with since. Stoicism came out strongly, with a general attitude that there was no point in dwelling on things that could not be controlled and therefore you had to think positively and carry on with life. Some respondents, when talking about their own experience replaced the term "depression" with "feeling down" as if this term was more acceptable to them.

"No, no I can't do with it. When I were a kid, my mother [...] she used to say "ohh pull yourself together" and that's what I do. I haven't a right lot of sympathy. You know, I used to say get on with it. That's what we all did, we got on with it. My mother had six of us and we got on with it, you know what I mean. So noooo. I might have been a bit down when I lost my husband. I'm just sorry I'm getting older and I can't do what I did 10 years since or yeah."

Mrs Twenty

Several participants linked this question to feeling down if they were stuck in and could not go out for a few days. For example, Mrs Twenty said "*Ohhh I couldn't do it if I were stuck in all day. I just pray to god that I'm alright and I can get about, you know. [...] I don't care about it (anxiety/depression) as long as I can (get out).*" There were also cases of narrow interpretation for this question with only depression or anxiety being mentioned for and not both.

A generational lack of acceptance of mental health and emotional issues was clear. Respondents commonly referred to the way they were brought up when discussing their stoicism and it was clear that, to many respondents, having issues with anxiety and depression were not socially acceptable. This was likely the reason for some respondents feeling this item was not relevant to their QoL.

Visual Analogue Scale (VAS)

Respondents tended to like the idea of the VAS as they had free range to place themselves wherever they wanted, rather than having to choose between specific responses. Response shift had a clear impact on the responses of some participants.

Several respondents discussed the fact that they did not expect to be in perfect health at their age and that at their age things could change very quickly. Direct evidence of response recalibration was seen when Mr Fourteen questioned the impact of expectations within his own answer saying that, despite being in objectively worse health now than when he was younger, his expectations were much lower now and therefore his valuation now might well be higher than the one he would have given at a younger age.

"I expect to go down a bit. You don't expect to stay the same active as you were 10 years ago and at the moment, I've put 75%. Well in a fortnight I could drop down to 50%, for all I know. Anything can change as you're getting older, so quick." Mr Two

"This scale thing is quite hard because, for example, I pretty well thought 75 but if I was 50 (years old), I wondered if I would put it quite a bit lower. So, it can be very misleading. I'm being critical... Its cliché, but everything is relative." Mr Fourteen

Response recalibration was suspected in the VAS responses of other participants, as despite some having substantial health issues, they rated themselves highly on the scale. Perhaps most notably, Mrs Four who is confined to a wheelchair rated herself as 90 on the VAS. Another example was, Mrs Eight, who lived with severe visual impairment, yet responded 75 on the VAS saying, *"normally I'm alright apart from my hand and my back."* She made no reference to her vision as if she viewed this as an entirely separate issue from her health. It was suspected that respondents may have interpreted perfect health as the best possible for their age and situation and made a relative assessment based on this, feeling that their health truly was good considering others their age. Finally, the researcher sometimes perceived a separation between respondents' view of their health in general and specific health issues which they were currently experiencing. It was sometimes felt that participants viewed questions about general health as only asking about whether they viewed themselves as healthy in general and that some participants did not include current health issues in this assessment.

Response issues identified during respondents' think-aloud completion of the SF-12v2 and subsequent verbal probing are presented in Table 40.

Measure as a whole

Participants had no issue with the length of the SF-12v2. However; some people struggled with the wording and layout. The questions are quite long, and participants often had to read them several times to understand fully what the question was asking, making comments such as "*Ummm right I've not followed that*" and starting again, like Mr Twelve. The layout of questions, which are often presented together in clusters, also caused some confusion with participants reading the long introduction to the cluster of questions without realising that the question was finished below and would therefore start answering before they had fully read the questions, for example Mr Twelve who said "*does your health now limit you in the following activities... what activities are we talking about... oh these here*" There was also a case in which a participant did not understand that more than one question was being asked in a cluster and would only answer one of the two or three questions.

Item 1 – General Health

All respondents thought their general health was an important aspect of their QoL and that this was an acceptable question to ask. When answering this question, respondents often mentioned either specific health conditions or general aches and pains. It was also common for patients to relate their health to their ability to achieve daily tasks, for example their ability to walk and get out and about or do their gardening. It was clear that participants' expectations had declined with age and that their view of their health was set relative to others of a similar age. Several respondents mentioned that their health could vary and decline quickly and that, due to their age, it did mean that tasks had to be done more slowly.

"My health in general, it's pretty good really. Oh god I can't grumble really. I can still walk I can still dig. Well I presume I can still dig. I haven't dug in a fortnight. I do all my own gardening. I can't do it as quick as I used to do it, but I can do it." Mr Two

Table 40 – SF-12v2 Response Issues

Each / represents a participant who experienced the corresponding response issue

	SF-12v2												
Response issue	Measure whole	General Health	Moderate acts	stairs	PR accomplish	PR limited	ER accomplish	ER careful	Pain	calm/peaceful	energy	downhearted/low	social activities
Practical Completion													
Length of measure													
Layout of measure	////												
Understanding													
odd wording			///							//	/		
difficult wording													
Recall													
Wrong time period		/									/		
Interpretation													
difficult interpretation													
wrong interpretation					/	/							
narrow interpretation					///	///	//	//					
Response Option Selection													
Format difficult													
Different answers for different aspects			///								/		/
Options partly applicable													
Irrelevant options													
missing options													
similar options													
order of options													
Inconsistent response		/	/	/	/	/	/	/			//	/	
Acceptability													
Item inappropriate													
Relevance/ Comprehensiveness													
Similar Question					//	//	/	/			/		
Item irrelevant				/			////	////			//		
Important aspects of QoL missing	//												

"I think it is an important question yes. I mean luckily for me, my health has been confirmed only a month ago, that I'm reasonably healthy for a man of my age like. I mean I don't expect to be Sebastian Coe (laugh) or Mo Farah like but I mean I could run for a bus. But I wouldn't want to like, but generally speaking, I feel generally healthy" Mr Twelve

There was evidence of inconsistent responding and ignoring the time frame of the question. Mrs Sixteen, a very active respondent who loved getting out and about, but who had sustained a recent hip fracture (more than 2 months before) and was not yet able to leave the house, responded that her health was excellent saying, *"Well I'm going to put excellent because my health in itself is... its only as I am at the moment due to that fall. Before that I'd got no problems whatsoever yep."* Responses such as this one suggest that when making subjective assessments of their general health, respondents sometimes distinguished between their view of themselves as generally healthy and specific health issues that they were currently experiencing. These issues, by being seen as separate to their health in general, were not included in the assessment of general health resulting in more positive responses than would be expected.

Items 2 and 3 – Moderate Activities and Stairs

There were some issues with the layout of this pair of questions, which affected respondents' interpretation of the questions and their responses. Several respondents had an issue with the wording of this question. The list of suggested activities in the moderate activities question includes moving a table, pushing a vacuum cleaner, bowling or playing golf. Several respondents sounded particularly surprised or amused at the inclusion of playing golf as they felt it was not something they would ever do. They also felt these activities required two very different levels of physical ability, which led to response option selection issues as different response options applied to the different examples of moderate activities provided. While many participants felt they could move a table or push a vacuum they would not be able, or would never try, to bowl or play golf. This led to issues of people not being sure whether to: ignore the suggestions altogether and interpret their version of moderate activities; to ignore the more vigorous examples, which they felt did not apply their life and respond according to the easier two activities; whether they should attempt to imagine how limited they would be in these more difficult activities and respond based

on that; or whether to provide a middle response over all the activities suggested. One participant, Mrs Six, proposed that maybe a more general example of housework as a moderate activity would be more appropriate. Although Mr Twelve did say his wife would laugh at the idea of him doing the vacuuming, so a less gendered phrase may be more appropriate.

"Moderate activities, moving a table I'm alright, pushing a vacuum cleaner I'm alright. Bowling or playing golf I just wouldn't do it... so is that being limited then? It just doesn't come up in my life. No not limited at all [...] But I've missed that bit out altogether – bowling and golf doesn't apply to me. But I mean everything else uhh I'm not limited at all, so I put that. [...] They should just put like housework instead of putting those. Put housework it would be better than that – golf. I can't do 10 pin bowling now and I've never played golf in my life, so I couldn't tell you (laughs)." Mrs Six

"Moderate activities such as moving a table, pushing a vacuum cleaner, bowling or playing golf. Well that's two different questions there, cos I mean I don't think I'd be able to play golf. Might be able to do bowling but I have arthritis in my neck which would prohibit me from doing those. I wouldn't call them two moderate activities. I would have thought they were a bit more sort of... I'd say limited a lot or limited a little because I mean obviously, I can push a vacuum or move a table. Soo.... I'll put a little [...] yeah, I think there are two questions there, because there is no comparison between running a vacuum and moving a table to playing golf on a 5-mile golf course. Like the two don't marry, do they?" Mr Twelve

What respondents felt was a moderate activity generally centred around being able to get out and about and being able to do their housework and gardening. For example, Mrs Fifteen responded *"Like dusting, washing. I can put the washer on and peg my clothes out and do my ironing. That's what I call moderate."* while Mr Two felt that *"Moderate activities is doing my garden. And walking down to the bottom (of the road) and back"*. All respondents felt that being able to do moderate activities such as these were relevant to their QoL and that this question was acceptable to ask.

Most respondents reported that they would be at least limited a little on stairs. Many comments about this question centred around adaptation with most respondents stating that they would struggle with several flights of stairs and look for a lift or escalator while in public. Most also reported that they felt less stable on stairs in their

older age and had had bannisters installed at home which they felt were essential to their ability to manage any stairs in the home.

"I have, this last year, fitted banisters on both the staircases and my wife says now I don't know how we'd have managed without this banister. It's one thing that she recognizes as being necessary. So, when I go up there, I'm holding the banister, whereas at one point I would have run up." Mr Twelve

"I've got a banister either side which I can grab hold of. So... I'm not sure how I would do with several flights. I don't think... I would be certainly limited quite a lot in fact these days even before I had my accident, I used to look for lifts and I used to love going up and down escalators, but I fell on one once and that put me off afterwards." Mrs Sixteen, 94

Inconsistent responding was seen, with one respondent, Mrs Six, saying that she would not be able to climb several flights of stairs but then responding limited a little. She also stated that this would never come up in her life as there would always be an alternative or she just wouldn't do it and therefore she felt this question was not relevant to her QoL.

Items 4 and 5 Physical Role Functioning

While some respondents interpreted these questions to include household tasks as well as activities outside the home, there were issues with narrow interpretation of these items as regular daily activities were often interpreted as self-care and housework only rather than broader elements of daily activities such as activities and being able to get out. Some respondents focussed on the "work" part of the question and therefore potentially felt that this may not be relevant to them. Respondents often felt that they accomplished less than they would like because ageing had caused them to slow down. This was often linked to frustration at not being able to achieve as much as they were used to.

"Well daily activities to me really is housework, cooking, housework I suppose. [...] I think that's valid as well, a little of the time. Because there are things that I now don't do, that I would have done... I'm trying to think... oh I know I wanted to go up into the top of the wardrobe and I've got a set of steps and I was wanting to change a handbag and I did go up the steps. I thought I'm not

waiting for one of the daughters to come, you can do that Mrs Three, but then it's a bit of nervousness more than anything." Mrs Three

"Well, all I've been doing is just getting my meals and washing up and that sort of thing you know since I've done this (broken hip). I hope in time that I will be doing more. I'm hoping to get out and about again, because as I say I never spent one day that I didn't go out somewhere." Mrs Sixteen

Being able to do their regular daily activities was obviously important to people and they had an obvious sense of pride in this and saw it as important to their independence. People often referred to adaptations they had had to make to be able to do their regular activities but emphasised the fact that they could still do them, even if they took much longer than they used to.

"I never, no matter what I'm doing, I never think I can't do that. For example, I bought some paint 2 weeks ago because these doors are just starting to show their knots, so I were going to do that, but I haven't got round to it yet because I thought I can't be bothered with that today like. But once I do, I'll probably only do an hour rather than go on until I'm absolutely wrecked. So, my health does restrict what I can do – well the fact that I'm old (laugh) [...] I expect it and because of that I purposely only do that amount. [...] I think it's important yeah. Important in so much as I've never had anyone do the work for me." Mr Twelve

While all participants felt that being able to do their regular daily activities was important to their QoL and acceptable to ask about, the similarity between the two physical role questions was noted by two participants who questioned whether both were needed.

"During the past 4 weeks how much time have you had.... Uhh daily activities – accomplished less than you would like... I would say a little of the time. Were limited in the kind of work or activities... I don't see any difference between them two... a little of the time." Mr Twelve

"It's the same sort of question isn't it. Some of the time it's all according to how you are day by day. Some days you can move mountains, some days you're knackered. Limited in the type of work and other activities, same thing. You do what you can." Mrs Eighteen

People did tend to focus on accomplishing less rather than being limited in the kind of activities they could do. Only one participant selected a different response between these two questions.

Items 6 and 7 Emotional Role Functioning

There was a reluctance amongst participants to recognise problems with anxiety or depression. Most respondents, while recognising that some people may suffer with emotional problems, such as anxiety and depression, said that this was not something that affected them or that they ever thought about. Most people referred to a stoic attitude, stating that they did not dwell on things they could not control and therefore they did not experience anxiety or depression.

"Well I don't get depressed. I'm not one that dwells on things an awful lot. [...] I wouldn't like to answer for everybody. I am not easily depressed but there are people who are. It's in their makeup somehow but I don't get easily depressed. I try to look on the bright side as much as I can anyway. I consider myself very lucky to have lived to this age with good health you know really and truly so umm its. No, I as I say some people do get depressed and very low and you can understand that too." Mrs Sixteen

There was an issue with inconsistent responding in relation to these questions. One participant who stated that she did not have any emotional problems then selected some of the time for accomplishing less than you would like. It was sometimes felt that participants lost the connection between accomplishing less/doing things less carefully and this being due to emotional problems. People would say they never had emotional problems and then they would answer that they accomplished less than they would like as if this were a separate issue. This may again have been due to the layout of the questionnaire with the long overarching question being presented slightly separately to the sub questions. Only one participant responded different levels between these two questions.

"Well you haven't to have emotional problems. You have to get on with it. Ummmm some of the time. Uhuh did work or activities less carefully than usual... hmmmm sticking to the middle of the road there so some of the time. [...] I think that's fair to ask but I'm not into emotion no. it doesn't worry me at all really. You have to get on with it, I'm sorry. I've always tried to be practical,

doing things and that comes down to practicality I'm afraid. You do accomplish less than you would like but..." Mrs Nine

The only circumstances under which participants spoke about having felt depressed was after the death of family members. Discussions of feelings of anxiety were most often related to past periods of financial problems or worries.

"Yeah, I mean, going back in life, I mean when I took this job at the newsagents I realized when I'd started doing it that I didn't have enough capital and I started taking on more than I could cope with. That's why I decided to sell the better house and come into here to give myself some capital and during that time... it took about 6 months to sell that house, I were a bit fretful. But once it were sorted, I were alright." Mr Twelve

However, any mention of experiencing these issues was always followed by assurance that this did not last long, and they no longer had any issues, as if anxiety and depression were viewed as only short-term reactions to big negative life events. Therefore, mostly people replied that while these questions could be relevant for some people, they didn't feel they were relevant to their QoL. All participants felt the pair of emotional role questions were acceptable to ask.

Item 8 Pain

All respondents agreed that pain could have a big impact on their ability to carry out their regular activities and therefore had a big impact on QoL. It was agreed that this was an important and acceptable question to ask.

"During the last 4 weeks how much did pain interfere with your normal work including both outside the home and housework. Well I haven't had any pain so that's just a little bit because, I do have arthritis but just lately its disappeared. Isn't that wonderful. Wonderful I don't know where it's gone to, it may come back (laughs). Ummm I've had 2 hip replacements and they've been absolutely wonderful. [...] Yeah. You don't even remember that you had pain and it was an awful pain, dreadful pain hip problems." Mrs Nine

"Well that's a question that's necessary isn't it. You need to ask that because there's varying types of health that a person might have. I'm assuming you're

interviewing people that are my age, not 30-year olds like. I know quite a lot of people my age who are worse than me like, so it's a fairly important question."

Mr Twelve

Pain was very important to the QoL of all respondents, as it substantially affected their ability to participate in social activities and their usual activities around the home. Participants often spoke about how pleased they were when issues of pain were resolved, and it was clear how their mood was lifted by this.

Item 9 Calm/Peaceful

Generally, respondents were happy to report feeling calm and peaceful at least most of the time. Being calm and peaceful was often related to thinking positively and not worrying about or dwelling on negative events or situations. For example, Mrs Nine said "*I do believe that it's all up here; well most of it is. It's no good getting into bed about things because there's very little you can do about it when you've got to your 80s*".

"Uhhh really the biggest part of time I think I have. I don't know if I'm one on my own, but I don't think that to worry about things gets you anywhere [...] I think you only, you stop yourself getting better if you start worrying forever. If I see an improvement, I am pleased about it, but I try not to worry about it."

Mrs Sixteen

Several male respondents however found the choice of the term calm and peaceful odd. One suggested that he would prefer to be asked whether he was content rather than calm and peaceful while Mr Seventeen seemed to feel that someone who was calm and peaceful all the time would be too relaxed and would not be able to achieve anything and so this was not necessarily a good thing, when he said "*You're joking (laughs). Who thinks these questions up... like smoking an opium pipe yeah. Uhhh I don't think anybody is always calm and peaceful you'd bloody fall over or do something wrong.*" However, nobody felt that this question was irrelevant or unacceptable.

Again, responses to this item demonstrate the importance that respondents placed on positive thinking and not dwelling on or worrying about negative events which they could not control. While some questioned the wording of this item, all respondents felt

it was important to feel calm in the sense that this meant they were not having to worry.

Item 10 Energy

People often struggled to choose a response option for this item because they felt that their energy levels varied a lot, both day to day and within a day. Several respondents, such as Mr Twelve, felt that the way the question was phrased was not appropriate or relevant to older adults because nobody their age would often have “a lot of energy” and that maybe rephrasing the question to ask how often someone had “enough energy” would be more suitable.

“Uhh not sure about that question there... a lot of energy. I’d say some of the time. [...] I think I just went medium in that one didn’t I. I mean some days I get out of bed and my necks aching and I think ohhhhhh [...] I think it’s an obvious answer question. I don’t think anyone at my age would say they have a lot of energy like. I’ve got enough. Enough yeah.” Mr Twelve

Some participants, for example Mrs Nine, also saw a distinction between the amount of mental and physical energy they had, and this led to a desire to choose different responses for different aspects of their answer, such as Mrs Nine who said “I don’t have a lot of energy no. I do in my head but not physically no. [...] If I had a bit more energy, I would do a lot more. If I could physically get around a bit more it would be better, but as long as I can drive, I can get there.”

Inconsistent responding and answering according to the wrong time period were seen with this question, with Mrs Sixteen responding that she had a lot of energy most of the time but saying *“I was always out and about (before hip fracture approximately 3 months before) anyway I did have a lot of energy.”*

While it was recognised that energy was required to achieve and participate in activities, it was something that participants felt varied a lot depending on the day and situation. The researcher felt that the phrasing “a lot of energy” caused issues with this question, as this was not something that necessarily reflected the current situation of respondents. This therefore encouraged inconsistent responding, referring to the past when they had more energy and feelings among participants that this question was not relevant to them.

Item 11 Downhearted/Low

Similar to the emotional role items, some people reacted to this question about being downhearted or low by restating the stoic attitude of needing to carry on, think positively and not dwell on or worry about things. Several respondents noted that this question was very similar to the emotional role questions.

"Being downhearted and low – no, no. [...] No, I don't go in for that. But it's relevant to some people and I'm sure they think it's a proper physical or mental problem, but for me it's not. If you can't do anything about it, it's not there. Or it is, but you ignore it." Mrs Nine

However, it did seem that more people were more willing to admit to sometimes feeling downhearted and low than depressed and anxious, as if these terms were more acceptable to them. For example, Mrs Three responded "*It is, and if you are down and feeling fed up and everything else that can ruin any day, can't it?*" Some people linked feeling downhearted and low to being stuck in and to loneliness.

"Ohhh, I feel downhearted sometimes. When you're on your own. I'm not as bad now with my neighbour coming in but at one time when she wasn't coming in. I used to go with her shopping, but she didn't call in. Now she calls in every day. It makes a big difference. A very big difference. People don't realise how much difference somebody calling in makes to a person on their own. It makes a hell of a difference. It does, it's the most important thing. Loneliness." Mr Two

It was agreed that the question was acceptable and, even if some people said that it was something that didn't affect them or that they didn't think about, it might be relevant to the QoL of other people and was therefore worth asking.

The use of the terms downhearted and low were felt to be more acceptable than anxious and depressed, which have been used in other items and measures. It was felt that some respondents interpreted downhearted and low as more temporary mood issues, by describing how it "could ruin your day", while anxiety and depression were interpreted more as longer-term mental issues. In this way, the use of the term downhearted and low was more socially acceptable as being low and in a bad mood

was more accepted than having long-term mental health issues which this generation are reluctant to recognise.

Item 12 Social Activities

When considering social activities, participants tended to focus on visiting friends and family rather than broader social activities such as social groups and activities that many participants had mentioned attending. Again, not experiencing emotional problems was often stressed and therefore if there were any issues with people's health that interfered with their social activities, participants emphasised that these were physical. For example, Mrs Nine emphasised "*I've put some of the time but its more the physical side rather than the emotional.*" This led to an issue in response option selection as participants wanted to select different responses according to the different elements of the question. People also mentioned that recent issues with not being able to get out to social activities were more do to with adverse weather conditions than issues with their own health.

WEMWBS

Response issues identified during respondents' think-aloud completion of the WEMWBS and subsequent verbal probing are presented in Table 41.

Measure as a whole

Most people found the layout and response options easy to understand and nobody had a problem with the length of the questionnaire. People sometimes struggled a bit more in their interpretation of the question and in selecting response options as these questions were more subjective and potentially covered concepts that they did not think about regularly. For example, Mr Nineteen, when asked how he found the measure to complete, responded "*Yes, yes some things you could interpret in different ways but apart from that reasonable yes*", which demonstrates the ambiguity which some respondents struggled with in some items in this measure.

Table 41 – WEMWBS Response Issues

Each / represents a participant who experienced the corresponding response issue

	WEMWBS														
Response issue	Measure Whole	Optimistic Future	Useful	Relaxed	Interested People	Energy	Deal Problems	Think Clearly	Feel good self	Close to People	Confident	Make up own mind	Loved	Interest New Things	Cheerful
Practical Completion															
Length of measure															
Layout of measure															
Understanding															
odd wording															
difficult wording															
Recall															
Wrong time period															
Interpretation															
difficult interpret									/						
wrong interpret					///				/						
narrow interpret					/				/				//		
Response Option Selection															
Format difficult															
Different answers different aspects						/									
Options partly applicable															
irrelevant options															
missing options															
similar options															
order of options															
Inconsistent response						///	//								
Acceptability															
Item inappropriate		//	//			//		/							
Relevance/ Comprehensiveness															
Similar Question															
Item irrelevant		///	//	/	/				/	//		///	/		
Important aspects of QoL missing	/														

Item 1 Optimistic about the future

Some participants questioned how relevant and appropriate this question was at their age, as they did not feel sure that there was much future left for them. For example, Mr One was amused by the question, saying “*(Laughs) I'm feeling optimistic about the future. How much future is there when you're 80? (laughs).*”. This often led to comments about what period of time to consider; tomorrow, a year, or more. Some people viewed it as having things to look forward to and mentioned holidays or events that were coming up.

“(Laughs) Feeling optimistic about the future (laugh) what future? Future is tomorrow or today; it's not beyond. You know, you wake up in the morning and you think oooo I'm alive again. But you don't think too long term, you know. At Christmas I think, well will I be here next Christmas... Feeling optimistic about the future (laughs). I don't think about the future really, I'll put rarely.” Mrs Nine

“Well at my age, you know what the inevitable is, so how can you be optimistic. This is why old folk get depressed. Because you're thinking what's there to look forward to. I'm just going to get more infirm, lose my mind. There isn't a lot to be optimistic about. But I like to think tomorrows another day, you know.”

Mrs Eighteen

Some said you had to feel optimistic as it was important to think positively. This idea was expressed by Mrs Sixteen when she commented “*I'm always optimistic (laugh) I think once you start being pessimistic you've had it... If you can't think positive, you think negative. It's no good at all.*” While some people viewed the likelihood of a future of gradual decline in functioning as something which people could find upsetting, some chose to interpret the question more positively.

Item 2 Useful

Most people felt that feeling useful was important to their QoL and could find examples of ways they felt useful to other people. Most commonly they felt useful to family or friends, or through helping people at social activities and clubs which they regularly attended. For example, Mrs Nine responded “*Yeah yeah, I think it's very relevant yeah. I think it's important you feel useful otherwise you just decline really. You just think what's my purpose of being here. I think it's important that you try to get involved with other things and try to help other people if you can.*”

Some were less positive about whether they still felt useful. This was usually frailer respondents, such as Mrs Thirteen, who had a lower level of independence and ability to help and responded “*Useful did you say? No, I don’t expect to be useful (laughs). I think I’ve done my job! I’m retired!*”.

“Not as much as I did because I used to look after the children, and I can’t now. Some of the time. [...] I think I’m alright with the other people here if they need something [...] I think I’m useful because the man next door. He hasn’t got a soul in the world and he’s really nice. He can’t walk very well. I’ve started taking him jelly and custard and tinned fruit and cake for his dinner. I think I’m doing a useful job for him.” Mrs Eight

For some respondents it was clear that their first reaction when asked about feeling useful was tied to roles they had performed in younger adulthood such as work and raising a family. These roles were clearly still important to them. Less frail respondents often tied the ways they felt useful to traditional family roles of women caring for the family and men doing practical jobs.

Item 3 Relaxed

Most participants reported feeling relaxed at least most of the time. People often said they felt much more relaxed now than when they were still working. Several participants, who said they could often be so relaxed they fell asleep, brought up the concept of being too relaxed.

“Well yeah, I do now because I don’t have anything to rile me now, so I always feel relaxed so. [Interviewer: good and is that important to you?] I think so, yes. I mean if you’re not relaxed you go to bed and you start thinking about all your problems and you don’t get no sleep and you’re tossing and turning. But that don’t happen to me now. I can sleep on a clothes line like (laughs).” Mr Twelve

The ability to feel relaxed was sometimes seen as a character trait. People often saw the opposite of relaxed as getting wound up and uptight about things and said that wasn’t the type of person they were, as worrying about things didn’t do anyone any good. For example, Mrs Sixteen responded “*I’m not one that gets uptight about things I never have been. I’ve just accepted them really.* [Interviewer: and do you think it’s

important to feel relaxed?] I do yeah, I do. If you're going to be uptight about things all the time your health is going to suffer, I think. It really is."

However, by contrast several participants saw themselves as the type of person to easily get worked up about things and therefore reported that they were rarely relaxed as it just wasn't in their nature. For example, Mrs Eighteen said, "*I've been feeling relaxed. Nooooo rarely. I never... I'm always (agitated/ wound up), as you can see. Rarely.*" She therefore did not feel this question was relevant for her, however everybody else stated that this was important, and no other issues were seen with this item.

Item 4 Interested in other people

Most participants felt interested in other people and felt this was an important aspect of their QoL. People mostly mentioned being interested in friends and family. Several of the more isolated and frail participants said this was no longer relevant to them so much anymore as many of their friends and family members had died and they could not go out and see the few still alive.

"Everybody really. I like people and I like meeting people and I think that is necessary as well to uhh... it gives you a further outlook on life too, being with other people and talking to other people. Like I talk to people on the bus and things like that which is great, hearing other people's views as well." Mrs Sixteen

"Not now no, because I've nobody really near. Because the people I was friendly with they've either died or... my friend is 78... no she's not she's 98. But she's in Workington in a nursing home there, to be near her son. So, there's no way I can go. My sister took me once but it's a long journey and I get tired. [Interviewer: did you think that was relevant?] Not really no. At one time, yes, when I was younger and able to go about and that." Mrs Ten

Several participants interpreted this question incorrectly and questioned whether feeling interested in other people meant being nosy, such as Mr Twelve who said, "*I'm not sure what's meant by that. Is that being a peeping tom, nosing out the window or (laughs) seeing what the neighbours are doing.*" This is an issue as these participants

viewed someone who was interested in other people all of the time as something negative, while this often is intended to be the most positive.

Item 5 Energy to spare

People mostly said that having energy was important as without it you could not do anything. Several participants noted the “to spare” and felt this made the question inappropriate for older adults, as it was unrealistic for an older person to often/always have energy to spare, such as Mrs Thirteen who responded “*No no no. That's mental isn't it... to spare! (Laughs)*” and Mr Twelve who said, “*Yeah I think it might be a question that shouldn't necessarily be asked that, because I think anybody of my age doesn't have energy to spare like*”. Similar to the SF-12v2 maybe the phrase “enough energy” would be more relevant to older respondents. The phrase “enough energy” was commonly used in respondent’s answers, such as the response of Mr One, who said “*I didn't know what to put, so I've put it some of the time. I've put it right in the middle because it, well... I've nowhere near as much energy as I used to have, I wouldn't say. I have enough energy, but I don't think there's a right lot left over at the end of the day (laughs).*”

Inconsistent responses were common to this question as people read the question and said that having energy to spare was unrealistic and not something they had, and then went on to answer more positively as if they were responding to “enough energy”. Again, some participants felt there was a difference in the amount of mental and physical energy they had, for example Mr Nineteen who said “*Energy to spare... whether that means physical energy or mental energy I don't know so I just put some of the time. I think I have got mental energy to spare and yes I suppose physical, to a certain point yes.*” This sometimes led to an issue in selecting a single response option. The amount of energy they had was something that participants felt varied a lot day to day.

Item 6 Dealing with problems well

Dealing with problems well was often linked with independence and was therefore considered something important to be able to do. For example, Mrs Eight said “*Yeah, I think all the time. I can cope. [Interviewer: and do you think that's an important part of your quality of life?] Yeah, I don't want to be dependent. Until I have to be.*”

Problems with the house were the most commonly mentioned. Often people noted that it took them longer to make decisions nowadays and described finding themselves worrying more about problems when they arose.

Some said that increasingly big decisions were discussed with family and others firmly stated how important it was to be able to deal with problems themselves. There were examples of inconsistent responding in relation to this question with several respondents selecting some of the time but also stating that they did not deal with problems themselves but relied on family for this.

"If I think there's out going to be wrong with the house, I get upset about that where at one time it wouldn't. [Interviewer: and do you think it's important to feel that you're dealing with problems well?] Well if there is a problem I don't. I wait until my daughter comes and I tell her, and she usually sorts things out."

Mrs Seven

"When you live on your own, you've got to deal with things. I say that, I've got a friend whose daughter lives up the road and she rings anything. You see I've been on my own a long time, so you get used to things. She's been a couple of years... well if out goes wrong she rings her daughter and says you'll have to come. Her daughter sorts all her bills for her. Well I do all that myself, you know what I mean. Cos I like to think I'm still capable of doing it. I know a lot of old folk don't do it. Their families do it for them." Mrs Eighteen

It was clear that dealing with problems was important to individuals' sense of independence. However, it was also one of the areas that they lost confidence in their abilities.

Item 7 thinking clearly

While people usually said they thought clearly most of the time and thought this was important, they had often noticed a decline with ageing. They often mentioned that it took them longer to make decisions or remember things in older age. For example, Mr One said "*Well at one time I would see a problem and decide what I was going to go, but now I sort of think well should I do this, or should I do that or... you know it takes me longer to decide. [...] But I can usually do it in the end yeah, so I think that one's alright.*" One participant questioned how appropriate this question was as, if the

aim of this question was to distinguish those with cognitive impairment, perhaps the individual themselves was not the best person to judge this.

"I still have it up here like, yeah. And if I'm thinking about doing something, I'll think about it yeah. [...] I mean there again I'm not sure if it's a correct question to ask because you'll know when you see a person. You'll think ohh he's into dementia or he's not." Mr Twelve

The ability to think clearly was linked to the idea of dementia by several participants, however this was usually in reference to themselves being lucky not to be in a state of cognitive decline and to still be thinking clearly. The ability to think clearly was valued by participants, who felt it was key to their ability to live and function independently.

Item 8 Feeling good about self

Some participants struggled to interpret this question. Some interpreted it as feeling good about themselves when they had made an effort in their appearance, while others thought about feeling confident and capable of achieving things independently. One participant, Mr Nineteen, felt that feeling good about yourself was linked to being conceited and therefore not necessarily a good thing. He responded "*Well I don't know. I don't feel good about myself necessarily. Some of the time on that. Yes, that's a bit self-centred isn't it really I think.*" One participant stated that he did not feel this question was relevant to his QoL as he did not think this way.

"Well I've ticked the often one, but I don't... I don't usually think about myself all that much, I just sort of carry on. I don't think ooo, I feel good today because I've done that or feel bad because I haven't done it things. No, it just carries on." Mr One

"To feel good about yourself, yeah I think so. You've got to feel fairly positive. To feel that you're capable of doing something." Mrs Nine

This question was felt to be important to QoL by those who interpreted it as feeling capable of achieving things independently. However, the ambiguity of the wording of this item left some respondents unsure what the item meant and therefore unsure of its relevance to them.

Item 9 Close to other people

Most people felt feeling close to other people was a relevant part of their QoL such as Mrs Nine who responded “*Yes, I've put often there, and I think so, because people are very good to me and I try to reciprocate and yeah I think it's important and I've felt quite close to people these past few weeks yeah*”. “Other people” were mostly interpreted as family but sometimes friends. Again frailer, more isolated participants thought this was less relevant to them and a better question for younger people, as they no longer had much, or any, contact with individuals they considered “close”. Men seemed more concerned about who they would class as “close”, as indicated by comments such as the two below.

“I've ticked often on that, but its uhh... I don't seem to get these sorts of well, I feel close to people, you know. I've got this circle of friends I know I can... if I have a problem, I know I can ring any of them up any time and they could ring me up, but I don't go around thinking oooo I feel close to so and so today or you know something like that.” Mr One

“Do I feel close to other people? Yes, one or two. Not en masse I don't think. But I do believe I'm liked. I like to be with people yes.” Mr Nineteen

Men would often refer to family and maybe one or two close friends as close but tended to emphasise where this line was drawn and say it was something they didn't really think about. Women did not seem concerned with making such distinctions and were happier to be more inclusive when considering those they felt close to.

Item 10 Confident

Interpretation of this item varied across respondents. Some people interpreted confidence in achieving things while others interpreted it as social confidence. In terms of achievement it was mostly linked to people's confidence in achieving things independently, from regular daily activities to bigger things such as going on holidays. Everyone felt it was relevant and acceptable to ask.

“Oh well I've ticked often for that yeah, cos I can deal with most things and uhh if I can't deal with it, I usually know a man that can (laughs).” Mr One

"Oh yeah, I can go into a room full of strangers it doesn't bother me one iota. This is what happened with this friend – she's come out a lot because I say just talk to them and she says ohh you can talk its different for you – I say just talk to people. But it doesn't bother me I'm not shy in any shape." Mrs Eighteen

Again, ambiguity was noted in this item, with interpretations felt to vary depending on what was important to them and their situation. Frailer participants mentioned more basic household tasks while healthier participants discussed social confidence or bigger activities such as holidays. Again, this may be evidence of response shift, as frailer participants adjust their benchmark as their ability to function independently declines.

Item 11 Able to make up own mind

Being able to make up their own mind was very important to participants and very closely linked to their independence. Everyone, no matter their living situation or level of frailty, felt this was important and relevant to their QoL, despite some of them having previously stated that they relied on family to help with problems.

"Uhhh yes, I've ticked some of the time on that because it takes me longer than it used to do to make up my mind what to do. But once I've made my mind up, I carry on and do it yeah." Mr One

"Oh yes yes, I like to make my own decisions. I don't like other people making them for me. I think that once you start doing that you lose a part of your life anyway really and truly. I think it's necessary that you are able to make your own. I know there are some people who can't and that is very very sad." Mrs Sixteen

An individual no longer being able to make up their mind was only considered an issue one might experience if they had dementia. For all other individuals it was considered possible and important to their QoL.

Item 12 Loved

Many respondents felt that feeling loved was important, in a similar way to feeling close to other people. When considering how it must feel not to feel loved Mrs Sixteen said "*it must be dreadful to feel that way. I don't know how I'd react to it really. It must be awful to feel like that. You feel lost, you feel as if nobody wants you do you and that is dreadful it certainly is.*" Feeling loved was clearly related to feeling connected to and having a role within the family or friendship groups, which gave people a role.

Again, frailer, more isolated respondents felt that this question was no longer relevant to them, in a similar way to feeling interested and close to other people. Mrs Ten described how feeling loved was no longer relevant to her as she no longer had family within an accessible distance. It was more basic regular social contact that was now important to her.

[Interviewer: ok umm feeling loved do you think that's relevant?] "Not really. There's a lot of elderly people who don't have anybody and they'll go to these nursing homes and that's them, nobody there. As long as I can have somebody coming in... I have a carer who comes twice a week to give me a bath. Otherwise anybody that comes, I can have a talk with them, like the window cleaner this morning. There's the tea man who comes every fortnight. Yeah so I can have a chat with them." Mrs Ten

However other not so frail respondents also questioned the relevance of this item, saying it was something they didn't think about and it was a better question for a younger person. This sentiment was expressed by Mr Twelve who said "*Uhhh that is a funny question at my age, feeling loved. [...] That sort of question wouldn't ever be in my mind. It might have been in my mind when I were 20, even 30, but not at my age now like.*" This was not because Mr Twelve was not in a relationship that he considered loving. He expressed that of course he loved his wife, but that "feeling loved" was not something he thought about or that concerned him. It seemed to the researcher as if he interpreted "feeling loved" as more associated with the feelings and worries people have early on in a relationship, whereas after decades of marriage this did not concern him and was not relevant.

Item 13 Interested in new things

Being interested in new things was often narrowly interpreted as asking about participants' interest in technology but some participants interpreted it more broadly to include current world events or new activities. People often felt that feeling interested in new things was important, so that they didn't become isolated, such as Mrs Nine who responded, "*Yeah you've got to be interested, otherwise you're just in a cocoon and you're too insular aren't you.*" Being interested in new things was a way to stay interested and connected with the world. A few respondents, particularly those who were frailer, felt this was not relevant to their QoL as they were the way they were now and were not going to change.

"Hmmmm, I'm not interested in computers or anything like that. I got my hob and a talking watch for my husband and a little clock. I'll try it if I think it's going to be useful. Some of the time, because I'm not one for wanting everything that everyone else has." Mrs Eight

"Ohhh I'm always interested in new things. What's going on around me. I've ticked often on that one. Yeah that's a good one. [Interviewer: and do you think that's relevant to your quality of life?] Yeah because If you're... it's, you know, sort of... finding out new things, doing new things, that keeps you going doesn't it." Mr One

Again, there was ambiguity in this item, seen in the differing responses given by participants. For those who interpreted the question as technology, it was seen as less relevant, as material things were not necessarily seen as important. However, those respondents who interpreted the item more broadly to include being interested in new activities, knowledge and world events tended to feel this question was more important as it gave them interests, topics to talk about and things to look forward to.

Item 14 Cheerful

Being cheerful was often linked to staying positive and carrying on. Most participants felt this was an important part of their QoL.

"Oh, I can't be miserable. To me life's too short. I know that there are times, and I have felt it, when I lost my brothers, which you do but they were also ones who liked to live life to the full. And they wouldn't want you to go on

grieving forever. They always used to say make the most of everyday and that's what I try to do." Mrs Sixteen

"Uhh well, I've ticked often on that. But its uhh yeah, because I don't do... I don't do feeling uncheerful (laughs) It uhh... I just carry on because I'm not uhh you know... it doesn't worry me." Mr One

Participants often linked their response to this item to the stoic attitude of looking on the brightside, staying positive and carrying on. It was also linked to making the most of life and recognising that they were lucky to be alive. These sometimes felt like repetitions of what participants considered to be socially accepted constructs, however these did fit well with the attitudes of the sample as a whole and the way they seemed to view their lives.

One of the frailer participants felt it was not relevant, as her life was what she expected at her age and that was it. This may suggest response shift again as this participant had lowered her expectations of life to such an extent that she no longer considered whether or not she was cheerful or in a good mood. However, the rest of the participants felt it was important to feel cheerful.

ONS-4

Response issues identified during respondents' think-aloud completion of the ONS-4 and subsequent verbal probing are presented in Table 42.

Measure as a whole

Nobody had an issue with the length of the measure, however issues were identified with the layout and format of response options. The layout of the measure caused issues for responses to the anxiety question, as many respondents failed to notice that the scale reverses here, with higher numbers not signalling better QoL as the other questions do. This is discussed further in the anxiety item section. One respondent said she found it difficult to settle on a number with so many options, while others preferred the more flexible format.

Table 42 – ONS-4 Response Issues

Each / represents a participant who experienced the corresponding response issue

	ONS-4				
	Measure Whole	Life Satisfaction	Worthwhile	Happy	Anxious
Response issue					
Practical Completion					
Length of measure					
Layout of measure	////				
Understanding					
odd wording					
difficult wording					
Recall					
Wrong time period			/		
Interpretation					
difficult interpretation				/	
wrong interpretation			////		
narrow interpretation					
Response Option Selection					
Format difficult	/				
Different answers for different aspects		/			
options partly applicable					
irrelevant options					
missing options					
similar options					
order of options					
Inconsistent response		/			
Acceptability					
Item inappropriate					
Relevance/ Comprehensiveness					
Similar Question				/	/
Item irrelevant		/	////	/	//
Important aspects of QoL missing	/				

"It's difficult to think of what the numbers are between themselves. I don't know why it's different that for me. I would never say in any event that I'm completely happy or completely... because I think I don't know. Maybe that's me, because I don't think I'm completely spot on. We're not we're human." Mrs Three

Some participants felt that their responses could vary substantially day to day and therefore struggled to choose a single answer that they felt reflected this. One participant suggested that it would be helpful to leave space for people to make further comments if they wished.

"It's not a question you can answer by just one answer. Depends what day it is. On the whole, I'm pretty... I could say 10 every day for all of them. But there's some days that would be totally different, so to me that isn't on this answer. You need more than the way you're asking it. ... I could be pretty good for a month and then get a day... and there's some days. I could go a few weeks where I'm not happy at all: loneliness. So, you'll have to work that out for yourself (Laughs)." Mrs Six

"The criticism I'd make of it is the lack of wriggle room, you know. Is there anything else you'd like to... it's always good, I think. Is there anything else you'd like to add, or would you like to make any further comment?" Mr Fourteen

This participant felt it was important to allow people to explain their responses and add anything which they felt may be relevant to an assessment of their wellbeing, suggesting that he did not feel these questions would provide a clear and comprehensive assessment of an individual's wellbeing.

Item 1 Life satisfaction

Health, social contact, being able to get out, being able to do what they wanted and not having to worry about money were commonly mentioned in people's evaluation of their satisfaction with their lives. The concept of life satisfaction was also often linked to happiness. Some said it could vary a lot depending on what was happening in their life or their mood.

"Well I suppose satisfied in that I have no pain umm that, we're back to money again, I don't have to worry about that. I have people uhh to call upon." Mrs Three

"How easy you are, you know. Have you any worries or anything. Umm we're alright, we're in a nice little house compared to what we were in. It were big. It were too big. It were costing a fortune to heat and light and everything. So, we're happy there. Overall, really everything. We can go out, we can stay in. We can do what we want." Mr Five

Again, response recalibration arose as respondents discussed shifting expectations and rating themselves relative to other people of their age in worse situations resulted in inconsistent responding in relation to this question. Some participants discussed how their expectations had declined, as they could now do less of what they had enjoyed when they were younger, but they still rated themselves highly despite this, as this is what they expected from old age. Several participants mentioned other people in worse situations to justify higher ratings of their own life satisfaction. Most participants agreed that life satisfaction was a relevant part of their QoL and acceptable to ask.

"Well it's what I expect when I'm 90. I don't expect to be 16 years old (Laughs). I used to work in a dementia ward, so I know what it's like to be worse than me. At least I know where I am." Mrs Thirteen

"Very satisfied generally. Yes, I am, I think so. I miss certain things that I did but you have to say, well tough you can't do it, you know. There's no point in sitting and moaning and feeling miserable about it. No, I think I'm quite happy."
Mr Nineteen

Again, in response to frustrations about activities which participants could no longer do, they adopted a stoic attitude and chose to focus on those which they could still enjoy.

Item 2 Worthwhile

Some respondents noted that it was unclear whether this question was asking whether the things they did were worthwhile to the community or to themselves. For

example, Mr Fourteen noted “*The problem is that it doesn’t say whether it’s worthwhile to the individual or to other people*” and Mr Nineteen responded “*Well I think they’re worthwhile to me yes. I don’t know if they’re worthwhile to society... I don’t know, but to me I think yes my life is worthwhile yes..*” It was clear from responses that both of these interpretations were being used and this had a substantial effect on how people answered the question. Those who were frailer and could do less tended to interpret the question as whether the things they did for themselves were worth doing, such as Mrs Ten who interpreted the question as “*being able to do my housework and cook my meals even though they are microwave and make meals when I feel like it*”. Conversely, those who were more able were more likely to think about doing things for others and feeling useful. This may indicate response recalibration as those who were less able chose to focus on more basic activities, lowering their internal standard. Some respondents looked back over their life to assess this question, so it was often not a current assessment.

Some people felt that saying they did not feel the things they did were worthwhile, was like saying they felt worthless and emphasised that they did not think like that.

“We’ve never felt you’re worthless like, have we. Some people seem to think that... but I’ve done nothing worthwhile or anything.... I’m worthless and stuff but we’ve never thought like that have we. We don’t really think about it... oh heck I’m feeling worthless, what can I do today. No, we never feel like that, because we’ve always got something to do haven’t we... There’s always something somewhere to be done. Yeah keeps us busy.” Mr Five

Many participants who interpreted the question in relation to the impact of their activities on others, stated that they did not base their decisions on what to do on whether it was “worthwhile” to others. They did not think about their activities in this way and did what they wanted to.

“Well for me, personally I don’t look at myself like that – that its worthwhile. But I know I’ve made other people happy and I’ve sorted out other people that needed a crack around the backside, so I don’t look at it like that. I know myself I’m happy with what I’ve done. What other people think don’t bother me. That’s their problem [...] I don’t think you ever look at yourself like that – is it worthwhile. What you do, you do and if it’s wrong you’ll soon know about it won’t you.” Mrs Six

A lot of people didn't feel this question was relevant, either because it was no longer relevant due to their age limiting them, or because they did not think about whether things that they did were worthwhile, they just did what they wanted to do, such as Mrs Ten, who responded "*No, I don't think. You do what you want, find your lot yeah.*" when asked whether this was a relevant element of her QoL.

Item 3 Happy

Everyone reported being happy, but many said this was something that could vary day to day. People generally agreed that it was a relevant part of their QoL and acceptable to ask.

"How happy are you feeling today? Well I'm alright, 7 again. I haven't been on my own, have I. My friend came, and you've come haven't you." Mrs Eleven

"Overall how happy are you feeling today. Um I'd say 7, you know. Not the least because I set myself some tasks today and so far, my tasks are being, well I hope, accomplished." Mr Fourteen

The assessment of happiness was commonly linked to social contact and the achievement of tasks and daily activities which provided respondents with a role and a sense of independence and achievement. One respondent questioned the need for both the happy and anxious questions as they thought if you were happy you would not be anxious and vice versa.

"They all mean something. Uhhhhh I should say the bottom two, how happy you're feeling today and how anxious you're feeling today, as they are you could do it in one question. If you're feeling very happy, you won't be feeling very anxious will you so that could just be one question." Mr Five

For this respondent the concepts of happiness and anxiety were seen as opposite ends on a continuum of current affect or mood. However, this was not mentioned by other respondents.

Item 4 Anxious

The layout of the ONS-4 caused issues for the anxiety question. Several respondents did not notice the direction of the scale changing for this question, leading to inconsistent responses and several more were surprised by it after starting to answer incorrectly and had to change their answer, for example Mr Five who said “*oh right, that’s wrong way round isn’t it that one... it’s a tricky one is that*”. Some respondents again referred to the stoic attitude of not being the type of person to worry and feel anxious and that they just carry on with life.

“I never feel anxious no. No don’t bother even though I’ve had a fall out with (child) I’m not anxious about it. It’ll either come right or it won’t. I’m not even anxious about what I said, about wanting more company. I wish I had a bit more company, but I’m not thinking ohhhhhh what am I going to do, how am I going to get through if I don’t get a bit more company. I’m not anxious about it, it’s just a thing I’d like. Otherwise id be crying all the time wouldn’t I. For goodness sake, get on with life (laughs).” Mrs Six

“I’m not anxious at all, that’s because you’re here. I’m not shooting you a line, it’s true. Its company you see. ... When we were first married, we always had to worry about money. You know, it was always a concern. We had a mortgage (my wife) was teaching and I was teaching, and it was always a concern. And when the children came along it was even more of a concern, but we managed. But now, on a teacher’s pension and my old age pension I’m not rolling in it, but I’m comfortable. I don’t have any anxiety about money at all.”

Mr Nineteen

Anxiety was most often mentioned in relation to financial concerns, but it was also often linked to a lack of social contact, family issues or problems with the home. However once again, if respondents mentioned one of these issues and having felt anxious about them, they were often quick to clarify that these feelings were only temporary. Most participants clearly interpreted anxiety as a temporary reaction to a negative life event rather than a long-term mental condition.

5.4.2.3 Comprehensiveness

Participants were asked at the end of each measure whether they thought that there was anything which was important to their QoL which was not covered by the measure. While the responses which were given for each measure are outlined measure by measure it was felt that they should be presented in a single section, as often what the participant mentioned for the first measure was also relevant for the second, but the researcher felt that they didn't want to repeat themselves. Equally sometimes they had gained more confidence by the end of the second measure and it was felt that they either had more ideas by the end of the interview or were more willing to voice them. Therefore, while ideas for additional dimensions for each measure are outlined according to the measure they were suggested for, they may be relevant and uncaptured by other measures more broadly and this can be kept in mind in this section.

EQ-5D-5L

Most participants felt that the EQ-5D-5L was comprehensive in covering their QoL. Three participants suggested additional dimensions which they felt would be relevant to the QoL of older people. These suggestions covered three main areas: relationships, social contact and loneliness; the way people were treated by others and feeling a burden; and being able to do what they wanted to do. People felt it was very important to be able to see their loved ones and friends when they wanted and to have regular social contact. Otherwise they felt they would be lonely, which people felt was a big problem amongst the elderly and something people felt it was very important to avoid.

"I've been lucky, since my husband died about (5-10) years since... I didn't know what I was going to do, and I saw these two friends walking down and they said we're going for a coffee, do you want to come, and I thought why not. I'm only going back to an empty house. So, I've gone with them all the time since then. And I say I don't know what I would have done, and she says we wouldn't have let you stop on your own – we'd have come for you... It's nice. I don't know what I would have done really. You could get lonely if you didn't couldn't you. You know if you didn't join in." Mrs Eleven

One participant noted that the way older people were treated was very important to their QoL. He pointed out that many older people he knew or met felt marginalised by society and felt they were a burden on both society and their families. Therefore, he felt a question about the way people felt they were treated could be of value. This could be a question around concepts such as dignity and respect and feeling involved, which again links to loneliness. Feeling secure about the future was also mentioned. This was both in terms of financial security as well as security about where you would end up in the future and whether services would be there if you needed them.

"What many older people in particular encounter is a feeling of being somewhat marginalized. What's interesting for example is that many of the old people will tell you at my work that they're not seen when they're shopping and so when someone calls on them, like the chiropodist, it really is the highlight of the week because loneliness as far as I can tell is one of the great difficulties." Mr Fourteen

"So many of my age group are very comfortable economically – some of my age group are very hard pressed economically and sadly although money is not supposed to be important I think worries about money in my experience and my work can be very profound not the least the anxiety of quote being a burden close of quotes and care in a unit and so there's always that feeling that just behind you there's the man with the pound signs written on him. In our area for example they're closing down council units so we are very worried about that because it's not a wealthy area umm and so I think probably the economic side uh is quite of concern." Mr Fourteen

The last potential area of improvement suggested for the EQ-5D-5L was whether people were able to do what they wanted. This was important to people in terms of control over their lives and independence and therefore a question surrounding these concepts could be relevant to include in a broader QoL measure for older adults.

SF-12v2

Two participants suggested aspects of QoL to be added to the SF-12v2 to improve its relevance to older people. These surrounded the concepts of loneliness and company, and coping. Mr Two, who lived alone, stated that "*I think loneliness is the*

most important question as far as I'm concerned." He felt that loneliness had a substantial negative impact on the lives of many older people and therefore a question around whether people had enough social contact and company was important. A question around coping and "*whether you get all the help you need*" (Mrs Nine) was also felt to be important to an older population.

WEMWBS

The addition of coping was suggested again by one of the frailer participants, Mrs Ten in relation to the WEMWBS. Again, this could cover aspects such as whether people have the support they need or whether they are able to cope with their life as it is now.

ONS-4

Mr Fourteen, who suggested several additions to the EQ-5D, felt similar suggestions were also relevant additions to the ONS-4. He mentioned again the concepts of security about the future and support. He reiterated that a big concern of older adults was security in terms of both financial security and whether the support they needed would be available for them when they needed it. Worries about these could have a big impact on the lives of older adults and he felt this was important to capture.

"Well again, I'm repeating myself, but obviously economic wellbeing and security and things like that which is very close, I think, to the heart of many who dread going into care. And the other thing is probably what could come into it somewhere is this idea of support. Umm not necessarily family support. But perhaps but I think it's very useful to know what sort of support is available." Mr Fourteen

As we can see similar additions were suggested by various participants for each of the measures and each suggestion would perhaps be appropriate for any one of the measures. These are important aspects to consider either as potential bolt on dimensions to existing QoL and wellbeing measures or as dimensions which could be important to include in a future measure of QoL and wellbeing relevant for older people.

5.4.2.4 Measure Preferences

In the final part of the interview participants were asked which of the two measures, if either, they preferred, and thought could do a better job of assessing their QoL. Of the twenty respondents eleven stated a preference for one of the measures over the other, while the remaining nine stated that they were both good. Interestingly of these nine who did not state a preference, four stated that they thought both should be used as they covered different areas of QoL. These four people all received one health measure and one wellbeing measure, suggesting that they recognised the different coverage of these measures and thought them both important. This was clear in the response of Mrs Sixteen, for example, who said “*No, I think they both cover a good spectrum anyway (WEMWBS SF-12v2). The questions I think are excellent. I think what one doesn't cover the other does. I think they're very good, absolutely.*” The remaining five participants simply said that they could not choose between them.

Of the eleven participants who stated a preference for one measure above another, there was no clear pattern of which measure was preferred over others. Peoples choices were dependent on both the type of question people preferred to answer and what people felt was more relevant to their situation. Several participants who received a health-related and a wellbeing measure said that they found the health-related measure easier to answer as the questions focussed more on their ability rather than more subjective feelings and concepts, which they didn't necessarily think about in their life.

“Yeah because I think that, as I say you don't really have to think about it too much. You see these (EQ-5D-5L) well ability again you don't have to think about it because yes, I can walk about... that's it. Whatever the question, it's perfectly clear which box to tick. With it done this way... when you get, it seems more difficult to decide on that type (WEMWBS) than on this type (EQ-5D-5L)”

Mr One

Several participants preferred other measures over the SF-12v2 because they found the SF-12v2 questions a bit confusing and felt the way that questions were asked on other measures was more to the point. For example, Mrs Nine said “*I think that this set of questions (WEMWBS) were more relevant – yeah yeah and more direct. I think that the others (SF-12v2) were a bit flannelly a bit waffley woolly.*” and Mrs Fifteen responded “*I thought that was better (EQ-5D-5L better than SF-12v2). The way it's put out for you. It's a lot easier to determine whether you put yes or no or... yeah.*”

5.5 Discussion

5.5.1 Summary of key findings

Older respondents in this study tended to define QoL in terms of four main themes; health, ability to carry out their usual activities, social contact and emotional functioning. Their health was central to their QoL, as their health determined their ability to undertake activities that they valued or enjoyed and their ability to access and participate in regular social interaction. Peoples' ability to carry out their usual activities was central to their definition of QoL as being able to do their regular daily activities in the home was key to their independence and being able to get out and about and engage in activities outside of the home was an important source of social contact. Independence was of great importance to participants. As their ability to undertake their usual activities declined, they discussed the need to adapt and eventually the need for support, either through networks of family and friends or more formally. However, the emphasis was always on those abilities maintained rather than areas where support was required.

Social contact with family, friends or people out in the community was of clear importance to all participants and loneliness was often discussed as a big problem amongst the elderly. The emotional impact of ageing was commonly discussed. The decline in health and ability to function independently, while somewhat expected by most participants, still caused frustration as they could no longer undertake activities or roles which they had enjoyed or valued and had to rely increasingly on other people for support. The speed with which their condition could decline also caused concern for the future. Participants often mentioned emotional strategies of stoicism and maintaining a positive outlook which enabled them to cope with the emotional struggles of ageing.

These themes closely reflect themes which have been found to be important to the QoL of older adults in the literature. For example, in-depth interviews used to determine the attributes of the ICECAP-O capabilities measure for older adults (Grewal, Lewis et al., 2006) identified six broad categories of factors which brought quality to the lives of older adults. These were activities, family and other relationships, health, home and surroundings, standards of living/wealth and religion/faith. These factors were important as they enabled social contact and attachment, enjoyment and pleasure, provided older adults with a sense of value associated with having a role and allowed them to feel secure physically, financially and through having a social

network (Grewal, Lewis et al., 2006). This very closely mirrors the themes discussed by older adults in this study and the way these themes impacted their QoL. Bulamu et al. also reported that older peoples' interpretation of QoL includes not only health but psychosocial and emotional wellbeing, independence, personal beliefs, material wellbeing and their environment in terms of its influence on their development and activity (Bulamu, Kaambwa et al., 2015). Older peoples' view of their QoL was found to be based on their ability to achieve those things and to participate in activities that they value and health was seen as a resource, which enabled their participation in activities of daily living and social interaction (Bulamu, Kaambwa et al., 2015, Grewal, Lewis et al., 2006, Milte, Walker et al., 2014). Similar findings were also identified in a study which investigated what was important to the QoL of older adults by asking them to rank states from the OPQoL-Brief and the ASCOT (Milte, Walker et al., 2014). The domain most commonly ranked first by respondents was health, followed by psychosocial and emotional wellbeing (phrased "I take life as it comes and make the best of things" and "I feel lucky compared to most people"), safety, dignity and independence/control.

The adoption of emotional strategies to deal with ageing and functional decline, such as stoicism and maintaining a positive outlook have also been noted in studies investigating the way older adults view their QoL and rate their health. Moser et al explored this in older adults with heart failure, a condition in which older people, despite worse prognosis and functioning, have often been found to report higher HRQoL than younger patients (Moser, Heo et al., 2013). This study found that younger respondents reported higher levels of anxiety and depression than the older group and that, for older people, the importance of a positive outlook was clear (Moser, Heo et al., 2013). The importance of a positive outlook when adjusting their perception of their QoL in the face of a chronic condition was also noted during a cognitive interviewing study of the EQ-5D and SF-36 (Robertson, Langston et al., 2009). It was felt that this stoic attitude and the importance of maintaining a positive outlook, despite problems, meant older adults were reluctant to signal issues to negatively phrased mental health items in PROMs studied in this thesis. This again could result in higher scores in older adults. Similarities with the literature provide support that the themes identified as important to the QoL of respondents in this study are generalizable to older adults more broadly.

Response issues were found for all measures which could threaten the validity of data obtained from PROMs. In the EQ-5D-5L there was narrow interpretation of the usual

activities domain, commonly interpreted as only asking about housework, and it was common for only one of the two concepts from the two double barrelled items to be mentioned. Inconsistent responding was seen throughout, particularly for mobility and the VAS, where response recalibration was suspected, with participants assessing their health in relation to what was expected for their age and what they saw in other members of their age group, rather than in relation to “best imaginable health”. Some respondents also had issues focussing on the time frame specified of “today”, which led to response option selection issues. It was clear that some respondents were averaging their answer over a longer time period as they felt, particularly for items such as mobility and pain, their state fluctuated day to day or according to specific situations or activities being undertaken. These issues have been identified in other studies, with the same narrow interpretation issues identified in older adults responding to the Dutch version of the EQ-5D-3L (van Leeuwen, Jansen et al., 2015) and inconsistent responding and time frame issues seen in older adults responding to the UK version of the EQ-5D-3L (Hulme, Long et al., 2004). The stoic attitude of participants led to questions surrounding the relevance of the anxiety/depression item. However, despite these issues, respondents liked the EQ-5D as they found the functional focus of the questions easy to respond to.

Issues of narrow interpretation were identified for the two pairs of role items from the SF-12v2, with “regular daily activities” commonly interpreted as household chores. Redundancy was also noted by participants within these two pairs of items. Participants found the examples of moderate activities provided in the question strange and unrealistic for an older population. This led to response option selection issues, where respondents felt unsure whether they should answer based on this list of examples and either state their limitations according to the most difficult activity or provide an average response based on the whole list; or whether they should answer based on what they felt was a moderate activity for them. These different strategies will impact their response and if different respondents choose different strategies, their answers are not necessarily comparable, as they are answering a question about different levels of activity. The relevance of the emotional role items was questioned due to the stoic attitude that ran through the interviews. Some participants also felt the ““have a lot of energy” item was irrelevant or inappropriate as having “a lot” of energy was unrealistic at their age. The layout of the SF-12v2 and length of the questions sometimes caused confusion and led some participants to prefer other, more concise measures.

Issues regarding the relevance and appropriateness of WEMWBS items to an older population were more widespread. These were particularly prominent for feeling optimistic about the future, useful, close to other people, loved and having energy to spare. Feeling optimistic about the future was felt to be irrelevant and inappropriate to some participants, who felt that at their age they could not be sure that there was a lot of future left as their health could decline very quickly and unexpectedly. Having energy “to spare” was also described as unrealistic at their age. While issues with the relevance and appropriateness of feeling optimistic and having energy to spare were identified by participants of varying frailty levels, the issues with feeling useful, loved and close to other people were concentrated in the frailest respondents. These participants shared that they no longer expected to feel useful as they could no longer do many things for themselves, let alone for others. However, it was clearly important to healthier participants to feel useful as it gave them a sense of purpose and it was clearly a source of pride. Those participants who questioned the relevance of feeling loved and close to other people were among the frailest and socially isolated respondents. Most of their social contacts had either passed away or lived too far away for them to be able to visit each other as they either struggled or were unable to leave the house. Therefore, they no longer felt that feeling close to people or loved was relevant for them and thought these questions were more suitable to younger adults. This could suggest response reprioritization as health and functional decline causes a shift in what is important or relevant to QoL. Ambiguity in the more subjective WEMWBS meant some respondents found them difficult to respond to. These respondents tended to prefer more practical, functioning focussed items.

Issues were also seen in the way older adults responded to and interpreted the ONS-4 personal wellbeing questions. Respondents interpreted the worthwhile item in different ways. For some it meant whether the things they did in their life were worthwhile to others in the community or to friends and family, while others interpreted it as whether the things that they did for themselves were worth doing or whether they could do what they wanted. The different interpretations, and the pattern in which they were seen, will cause issues when analysing responses to this question. It tended to be the frailest respondents who interpreted this question as whether they could do things that were useful to themselves. These were often basic tasks; such as household chores, preparing meals and doing things they enjoyed, such as reading. It was felt that this interpretation was largely due to response recalibration, as they no longer felt it was possible to be useful to others and therefore did not consider this in their answer. By focussing on basic tasks which they could still achieve they could still

rate themselves fairly highly. However, healthier members of the sample were much more likely to provide an answer based on what they did for others. This means that similar responses to this question may not be comparable as they may be answering different questions. Some participants felt that the worthwhile question was not relevant to their wellbeing, either because they did not think about the things that they did in terms of whether they were worthwhile or not, or because they no longer felt that they could be of use to others.

The stoic attitude of participants resulted in some participants feeling the ONS-4 anxiety item was not relevant to their wellbeing. The response scale of the anxiety item also caused problems. Participants often failed to notice that the end of the scale that indicated the most positive response, was reversed for this item, as it was the only negatively worded item in the scale. Therefore, participants often gave invalid responses, stating verbally that they were not anxious, but selecting a numbered response towards the end of the scale labelled “completely” anxious. This issue has been seen in previous content validation work for the ONS-4 and yet the response scale remains unchanged (Ralph, Palmer et al., 2011).

Response shift was seen in the ways participants assessed their own health and QoL generally and responded to items on the PROMs. Most respondents mentioned that their expectations of their health had declined substantially from earlier adulthood. Despite having ongoing health issues, which often impacted on their daily life, they continued to view their health fairly positively, as they viewed it relative to others they knew of a similar age who were in a worse state. Response recalibration was seen in participants' responses to a variety of items, including broad global assessments of health and life satisfaction as well as more specific items asking about physical functioning and ability to carry out activities. Where response recalibration is seen there is a risk that, in comparison with a younger person in the same health state, the older respondent would rate this same state much higher.

These response mechanisms have been noted in older adults' responses to PROMs in the literature. Several studies have identified that older adults' responses are affected by reduced expectations of their health and QoL in old age (Mallinson, 2002, Moser, Heo et al., 2013). Semi-structured interviews in the Moser et al. study of health failure patients, found that a major factor behind the better HRQoL scores among older respondents was the fact that their expectations were simply lower. While both age groups acknowledged that heart failure had a negative impact on their HRQoL

and what they could do, older respondents said that their HRQoL exceeded their personal expectations given their age or the fact that they could be dead, while younger adults had expected their HRQoL to remain higher for much longer and were therefore disappointed not to be able to perform activities and roles that they expected to (Moser, Heo et al., 2013).

The phenomenon that older adults judge their health and QoL relative to those around them of a similar age and situation who were worse off has also been seen in the literature. The Moser et al. study reported that older adults put their declining functional status in perspective by comparing themselves to others who they judged to be in a worse condition (Moser, Heo et al., 2013). A study by Ubel et al, also found that people's evaluation of their physical health was a relative process, in which individuals compared their physical health in relation to others in the same age group (Ubel, Jankovic et al., 2005). This issue was also found in qualitative content validation interviews of the SF-36v1 items in older respondents (Mallinson, 2002) and cognitive interviews examining the reference frames older adults with a chronic metabolic bone disorder used when responding to the SF-36 and EQ-5D (Robertson, Langston et al., 2009). It was also suspected to be the cause of unexpectedly high VAS scores from older adults in a study by Hulme et al., in which interviewers felt that older adults were interpreting the "best health imaginable" anchor as the best the respondent could expect to become, given their age and situation rather than perfect health (Hulme, Long et al., 2004). Again, this phenomenon shifts the benchmark for what constitutes good health downwards in older adults and may be another potential source of positive responding, which could result in higher expected scores in older adults.

Response recalibration, reprioritisation and reconceptualization was also seen in frailler participants interviewed. It was clear that the ability of the frailest individuals in the sample to complete daily activities independently had declined. Independence was still important to them and so they focussed their responses on basic tasks which they could still achieve independently. This recalibration allowed them to still answer fairly positively. However, these participants went further into the other elements of response shift. When faced with questions which required a higher level of functioning, such as feeling useful, these participants responded that these questions were no longer relevant to their QoL as they no longer expected to feel useful to others. Feeling useful was of clear importance to the more able participants, which suggests that once functioning declined and participants felt they could no longer

achieve things that had been important to QoL, they shifted their priority to other aspects of QoL. Another example of reprioritisation from frailler respondents was the increasing importance of their home and local environment. As their health and physical functioning declined, the layout of their home and the accessibility of local amenities increasingly determined how well they could carry out activities in the home independently and how well they could get out and participate in activities and social interaction.

Several participants suggested additional dimensions, which could improve the comprehensiveness of each of the measures for older adults. These were commonly applicable across the four measures discussed. These covered aspects such as: relationships, social contact and loneliness; coping and support; security about the future; control/whether people could do what they want when they want; and the way people were treated. These all appear reasonable suggestions in terms of relevance to the QoL of older people as these were all concepts that were often brought up when respondents discussed what was important to their QoL in the first section of the interviews.

These suggested additional domains align closely with the domains of other measures developed either specifically for older people (ICECAP-O) or with their qualitative input (ASCOT). The ICECAP-O domain of love and friendship aligns with the suggestion of a question around relationships, social contact and loneliness, while the thinking about the future ICECAP-O item corresponds with the suggested addition of the concept of security about the future. The concepts of control/whether people could do what they want when they want, somewhat align with the ICECAP-O items of doing things that make you feel valued, enjoyment and pleasure and independence. On the ASCOT; the social participation item corresponds to relationships, social contact and loneliness; the control and occupation items align with the concepts of control/whether people could do what they want when they want; the dignity item corresponds with the suggestion of an additional item asking about the way people are treated and different aspects of coping and support are covered by the measure as whole. This suggests that these items are indeed important elements of the QoL of older adults.

Peoples' measure preference was dependent on both the type of question people preferred to answer and what people felt was more relevant to their situation. Several participants who received a health and a wellbeing measure said that they found the

health measure easier to answer as the questions focussed more on their ability rather than more subjective feelings and concepts, which they did not necessarily think about in their life. Some participants who received a health and a wellbeing questionnaire recognised the different coverage of these measures and stated that they felt both were important.

These findings have important implications for the use of these measures in older adults. Content validity is argued to be the most important element of measurement performance (COSMIN Group, 2018), as the validity of the data received from questionnaires is dependent on whether the questions and response options are understood by the respondent, whether those questions are relevant to the concept that the instrument aims to measure and whether the important aspects of that concept are comprehensively captured by the instrument. The layout and style of question caused confusion on the SF-12v2 and ONS-4, which impacted the validity of responses received. The relevance of at least one item was questioned by some respondents on every measure, however this issue was more widespread on the WEMWBS, ONS-4 and SF-12v2 than it was on the EQ-5D-5L. The more subjective wellbeing items and negatively worded mental health items were more commonly felt to be irrelevant to older participants as they reported that they did not often think about their life in this way. This was particularly true for frailer older adults, who felt very few of the subjective items of the WEMWBS and ONS-4 were relevant to their life anymore, as their functional abilities had declined and therefore basic functioning were more their focus, rather than broader elements of QoL connected to having a role and purpose and social connection. It was felt that participants often found the more concise and practical functioning focussed EQ-5D-5L items easier to answer and relevant to their daily life, with the exception of anxiety/depression. These findings would therefore suggest that, of the four measures tested, the EQ-5D-5L may be the best starting point for measuring the effectiveness of health and social care interventions for older adults. However, participants did not feel that this measure comprehensively covered what was important to their QoL. Therefore, this measure may need to be adapted to make it appropriate for a comprehensive assessment of the QoL of older adults. This could be achieved through permanent adaptation of the measure or through the addition of bolt-on dimensions which could be used in older respondents. The implications of the thesis as a whole on the use of these measures to evaluate health and social care services aimed at older adults is considered further in the next chapter.

5.5.2 Limitations

At times, it felt as though respondents were prone to positive answering both within their responses to the questionnaires themselves and to questions about how they found the questionnaires in terms of relevance, acceptability and comprehensiveness. Although it was made clear in the introduction that these were not surveys that had been developed by the researcher and that all opinions about the questionnaires, both positive and negative were welcome, as this was the aim of the study, it was clear that some respondents remained reluctant to criticise.

While recruiting from an existing cohort made recruitment much easier, it may limit the generalisability of results. While a wide range of individuals were involved in the cohort, including people with a wide range of conditions, frailty status and living situation, the cohort did not include people living in care homes. Therefore, the results of this study may not be generalisable to a care home population. While non-white British individuals living in the Bradford and Leeds area were invited to take part in the study, none responded, and the resulting sample were all White British. The sample also generally, with a few exceptions, reported a lack of formal educational qualifications. However, for this age group, this may be fairly typical and while a lack of formal qualifications was reported a range of employment backgrounds from manual labour, nursing, teaching and engineering were reported. The sample did include a range of community living situations and frailty scores. These findings may therefore be quite broadly generalisable to an older community dwelling White British population however, content validity in older people from non-white cultural backgrounds and those living in care homes may differ.

Additionally, peoples' participation in the cohort involved them being visited every six months for five years. At these visits they were asked a wide range of questions about their health, QoL and ability to function including the EQ-5D-3L and the SF-36. The fact that they were used to being asked not only these questionnaires specifically but more broadly, questions about their health and QoL may have affected their responses to questions about the validity of these measures. They may be more accepting of questionnaires such as these than elderly people who have never seen these questionnaires before. On the other hand, the fact that they have experience in thinking about their health and QoL at these regular assessments may mean that they find it easier to form opinions about these concepts.

The process of identifying and classifying response issues is quite subjective. A large part of identifying response issues is assessing whether participants understood terms or questions in the way that measure developers intended. While a concept guide, outlining the intended meaning of terms and items, has been developed and published by the Euroqol group for the EQ-5D (Brooks, Rabin et al., 2003), evidence on developer's definitions of included concepts are much more limited for the other measures. Definitions for a limited number of concepts from the WEMWBS was found on their website in a document aimed at outlining and resolving common issues identified in translating the WEMWBS (WEMWBS Research Team). While this document did not cover all included concepts, it did provide some helpful clarification around the intended meaning of some of the less obvious terms included in the WEMWBS, however the intended meaning behind some terms remains unclear. For the SF-12 (and SF-36) and ONS-4 no concept guides could be found which clarified the exact intended interpretation of items. For these two measures, as well as the concepts of the WEMWBS not covered by the document identified, the researcher had to rely on previous published development and validation literature surrounding the measures in order to attempt to deduce the intended meaning of developers. There may be increased subjectivity and bias in resulting response issues relating to narrow or incorrect interpretation of respondents to these measures.

Subjectivity in identifying response issues was also felt to be a particular problem when looking for inconsistent responding. For questions about functional ability it was relatively easy to identify where participants had rated themselves substantially more positively than either they had stated elsewhere in the interview or the interviewer had observed from their behaviour during their meeting. However, for more subjective questions, such as happiness and life satisfaction, it is much more difficult to assess whether the response given is inconsistent with the respondent's true situation as the assessment is so broad and subjective. Therefore, it was much more difficult to assess inconsistent responding due to ambiguity or response shift on the two measures of subjective wellbeing than on the more functioning focussed HRQoL measures. This may have led to bias in this aspect of results.

5.6 Conclusion

This chapter aimed to investigate the content validity of four existing QoL and wellbeing measures in older adults. First, what was important to the QoL of older adults was examined to provide context to the way they assessed their QoL. Then cognitive interviewing techniques were utilised to examine the response processes older adults adopted when responding to the selected measures of health and wellbeing. Any response issues which could threaten the validity of data obtained from these PROMs and any comparisons or decisions based on this data were identified and examined.

Older adults' conception of what was important to their QoL centred around their health, their ability to carry out their usual activities, social contact and emotional wellbeing. Health was central to their QoL as it determined their ability to do their usual activities, get out and about and do what they wanted. Peoples' ability to carry out their usual activities was of utmost importance to their QoL, as being able to do their regular daily activities within the home was key to their independence. Being able to get out and about and engage in activities outside of the home was an important source of pleasure and facilitated social contact. Social contact with family, friends or people out in the community was of clear importance to all respondents and loneliness was often discussed as a big problem amongst the elderly. Emotional strategies of stoicism and maintaining a positive outlook were adopted by many participants to combat the frustrations of ageing.

Response issues were found for all measures which could threaten the validity of data obtained from PROMs. Issues with interpretation were found for some items from all measures, with respondents commonly narrowly interpreting questions about usual activities to mean solely housework. Some of the more ambiguous subjective questions were also misinterpreted by participants. For example, the worthwhile question on the ONS-4 was commonly interpreted in different ways, to mean either worthwhile to themselves or to society, whereas on the WEMWBS being interested in other people was commonly interpreted as being nosey. Differing interpretations have a substantial impact on responses and call into question the validity of comparing responses between individuals and groups. Participants most frequently questioned the relevance to their QoL of negatively phrased mental health and emotional items on the EQ-5D, SF-12v2 and ONS-4 and questions about feeling optimistic about the future and loved on the WEMWBS. The SF-12v2 and ONS-4 layouts and questions caused some confusion.

Response shift impacted the way in which older adults assessed their health and QoL and the way they responded to items on the PROMs. Response calibration was commonly seen, as older adults lowered the benchmark for what constituted good health, QoL or wellbeing by having generally lower expectations and assessing themselves relative to others of a similar age who were worse off. Response reprioritisation was also seen amongst the frailest respondents, who tended to feel that some items were no longer relevant to their QoL, as they could no longer perform many of the activities and roles associated with the items and had limited access to close social contact. These elements of response shift will impact scores and limit comparability between scores obtained from older and younger respondents.

In the next chapter, these results are compared to the results of the previous chapter, which used psychometric methods to investigate the construct validity of these measures in older adults. The implications of these findings for the use and development of QoL measures in older adults is also discussed.

Chapter 6

Discussion

6.1 Introduction

The aim of this thesis was to investigate the psychometric performance of existing PROMs in measuring the health, QoL and wellbeing of older adults. This information is useful to explore whether these PROMs are suitable for use in the economic evaluation of health and social care services aimed at older adults.

Three objectives were set in order to achieve this thesis aim. First, it was necessary to systematically review the existing literature on the psychometric performance of the chosen measures. This was required as the existing literature provides an important source of evidence on the psychometric performance of the chosen measures in older adults. It was also important to identify gaps in the literature, where there was little or no evidence of the psychometric performance of the chosen measures in older adults. This information was used to guide the focus of this thesis.

Second, given the findings of the systematic review summarised in the next section, it was important to explore the structural and construct validity, acceptability and internal consistency of the EQ-5D-5L, SF-12v2, ASCOT, WEMWBS and ONS-4 in assessing the health, QoL and wellbeing of older adults. These psychometric properties were investigated using IRT as these methods offer important additional information over CTT methods. IRT methods enable the detailed examination of the performance of item levels, the examination of DiF and estimation of internal consistency and precision of measurement at each level of underlying trait. Yet, they were found to have been underused in the existing literature identified in the systematic review.

Third, given the lack of content validation of these measures in older adults from a UK perspective, it was necessary to examine the content validity of the EQ-5D-5L, SF-12v2, ONS-4 and WEMWBS in assessing the health, QoL and wellbeing of older adults. Cognitive interviews were used: to gain an in depth understanding of

respondents' interpretation of items and response options; to examine whether they felt the items were relevant to their QoL and appropriate to ask; and to assess whether the measures comprehensively covered what was important to their QoL. This study not only added to evidence on the content validity of these measures in older adults, but also helped to further explore and examine the findings identified in the IRT study.

This final chapter presents the key findings of this thesis. It discusses the integrated results of the studies within the thesis, analysing whether these complement and contrast each other. It presents the contributions of the thesis to the existing knowledge and the implications and recommendations it provides for the evaluation of health and social care interventions aimed at older adults and the development of PROMs capable of comprehensively measuring the QoL of older adults. Lastly, it discusses the limitations of this thesis, and outlines recommendations for future research in this area of study.

6.2 Integrated key findings

This thesis and the studies it includes, provide valuable findings which contribute to the literature on this topic. While the individual studies within the thesis each offer important findings about the performance of the included measures when considered as separate studies, they also provide additional value when considered together. One of the great advantages of carrying out both the statistical psychometric testing and the qualitative cognitive testing is that the qualitative study can provide deeper insight into the way respondents interpret and react to items and measures. This can help explain, as well as validate, findings identified in the IRT study.

The key findings of the individual studies have already been discussed in the discussion sections of Chapters 4 and 5. In this integrated findings section, the key findings from each study within this research program are triangulated. This draws on where results of the two studies complement and contrast each other. First the measure specific integrated findings regarding the psychometric performance of each of the included measures will be discussed, before the broader findings surrounding the way older adults think about QoL and the ways in which the comprehensiveness of the included measures could be improved for older adults are considered. This provides a rich overview of the findings that can be taken from the thesis as a whole,

which is a good starting point from which implications and recommendations for future practice can be considered later in this chapter.

Measure specific findings

EQ-5D-5L

The multiple-group IRT model of the EQ-5D-5L found that mobility and usual activities provided the most information and were therefore the two items most closely related to the health of older adults, while anxiety/depression was least closely related to the health of older adults. These findings were mirrored in the qualitative study by the fact that being able to get out and about, carry out usual activities and social participation were central to the QoL of older adults and health was often viewed as the mechanism by which these functionings were maintained. Conversely, the finding that the anxiety/depression item was least closely related to the health of older adults in the IRT analysis reflects the stoic attitude of participants in the content validation study. Participants often expressed that negatively worded mental health items were not relevant to their QoL as it did not help them to dwell on things that they could not control, and it was important to carry on with life.

The investigation of DiF in the EQ-5D-5L found that older adults with the same level of health as younger people were more likely to respond higher (signalling better health) to pain/discomfort and anxiety/depression, with small effect sizes identified for these items. A potential explanation for the fact that older adults are more likely to respond higher to the anxiety/depression item can be drawn from the stoic attitude identified in the content validation study. This could have resulted in a reluctance among older adults to report problems on negatively worded mental health items, while in younger generations, there is increasing awareness and acceptance of mental health issues, which may mean they feel more comfortable recognising and signalling experience with such issues.

Another important finding from the IRT study was the presence of a substantial ceiling effect for the EQ-5D-5L, which resulted in reduced internal consistency for those individuals above 0.7 SDs above the mean level of health. The finding that DiF is resulting in older adults rating their health higher than younger adults with the same underlying health is likely to be contributing to this ceiling issue. Evidence of response shift found in the cognitive interviewing study, which caused older adults to lower their benchmark of what they consider to be a good health state, is a likely source of the

general DiF seen across the two health measures, which resulted in up to approximately 10% higher expected scores in older adults in the EQ-5D-5L and SF-12v2 in the IRT study. This is an important consideration as ceiling effects reduce the ability of a measure to precisely distinguish the level of health of respondents at the top end of the score range of the measure. It is likely that DiF is adding to this issue in older adults, by inflating their scores.

SF-12v2

In the IRT analysis, item redundancy was suggested within the multi-item scales of the SF-12v2. Redundancy within the two pairs of role items was noted by several participants in the content validation study, who felt that the items within each pair covered the same thing.

The role items provided most information about the physical and mental health of both age groups in the IRT study, suggesting they were most closely related to the health of respondents. The moderate activities and social activities items were also found to be closely related to the physical health of respondents aged 65+, with the next highest discrimination parameters. Again, this mirrors the findings from the content validation study surrounding what is important to the QoL of older adults, where being able to get out and about, carry out usual activities and have regular social contact were found to be central to the QoL of respondents. The relevance of the energy item was questioned in the content validation study, as participants felt the phrasing “have a lot of energy” was unrealistic at their age. This could also explain why this item was found to be among the least closely related to the physical health of older adults in the SF-12v2 IRT analysis.

In the content validation study, issues of narrow interpretation were identified for the two pairs of role items from the SF-12v2, with “regular daily activities” often interpreted as household chores. This is similar to the finding from the EQ-5D, where usual activities were commonly interpreted as housework. Participants questioned the relevance of the SF-12v2 emotional role items, which again could be explained by the stoic attitude towards negative mental health and emotional problems which resulted in a reluctance to acknowledge the relevance of negatively phrased emotional questions in the EQ-5D, ONS-4 and SF-12v2.

The DiF analysis of the SF-12v2 revealed that older adults were more likely to respond higher and signal better health than younger adults on a range of items including general health, physical role items, pain, energy, social activities and the mental health item pair. The finding of older adults being more likely to respond higher to questions about pain matches the finding for the pain/discomfort item on the EQ-5D-5L. An understanding of some of these DiF effects can be drawn from the content validity study. Again, the stoic attitude among older adults, could have resulted in a reluctance to report problems on the mental health items in older adults, while younger adults may be less reluctant. However, in the IRT analysis DiF was not identified in the emotional role item pair, as may have been expected from the stoic attitude towards mental health. This unexpected finding may be due to respondents reacting differently to the emotional role items than they do to the mental health scale. It is possible that the phrasing of the emotional role items may not cause the same stoic response in older respondents. It is also possible that the suspected inconsistent responding to the emotional role items, which was discussed in the content validity results, has an impact on the data obtained from these questions. It was often felt during the content validation study that, due to the layout and lengthy wording of this item pair, older respondents sometimes lost the connection between accomplishing less/doing things less carefully and this being due to emotional problems such as anxiety and depression. People would say they never had emotional problems and then they would answer that they accomplished less than they would like, as if this were a separate issue. This may explain the discrepancy between the results of the two studies and is something which should be explored further.

Examples of response shift identified in the content validity study, such as lower expectations of health in old age and the judging of their own health in relation to others of a similar age who were worse off, may also explain the tendency for older adults to receive higher expected scores to health-related items. It is likely that these frames of reference for older adults' responses would result in a lower benchmark for good health than the benchmark that a younger adult would adopt. It is therefore likely that older adults would rate the same health state higher than a younger individual which matches the tendency across the two health measures for older adults to receive higher expected scores than younger adults with the same underlying level of health.

ASCOT

For the ASCOT, occupation (worded as being able to do things you value and enjoy) was found to be closest related to the SCRQoL of older adults, followed by control and social participation in the IRT study. These concepts mirror the importance of being able to get out and about, carry out usual activities independently and participate in regular social contact seen in the content validation study.

DiF analysis for the ASCOT showed that the occupation and control items had the biggest DiF impact, with older adults expected to score lower (signalling worse QoL) on these items than younger adults with the same underlying level of SCRQoL. This may be due to the fact that control over daily life and being able to do the things that they value and enjoy were of central importance to the QoL of older adults and they felt the loss of ability to achieve these things keenly. However, it is difficult to say whether it is expected that they would feel this loss in ability more than younger social care users without conducting a similar qualitative study in this specific and distinct group of younger adults.

WEMWBS

In the IRT study, feeling useful provided the most information about the wellbeing of older adults in the external factor, while feeling optimistic about the future provided the least. Again, this echoes the findings from the content validation study. Respondents questioned the relevance to older adults of feeling optimistic about the future, as they did not feel they necessarily had much future left at their age and, as their health could decline quickly and unexpectedly, they did not think or plan far ahead. However, feeling useful to others was clearly linked to participants' sense of independence and feeling involved in family or community life and participants often expressed that feeling useful was important as it gave them a purpose. In the internal factor, feeling confident, good about oneself and cheerful provided were most closely related to the internal wellbeing of older adults, while having energy to spare was least closely related. Again, having energy to spare was one of the items that was most often questioned in terms of its relevance to older adults. This may have been due to the phrasing as some participants felt that having energy "to spare" was unrealistic at their age, which reflects the similar finding for the SF-12v2 energy item that also asked participants whether they had "a lot" of energy. On the other hand, feeling confident was often linked to respondents' ability to carry out usual activities and was clearly

linked to their independence, both of which were very important to their QoL, while feeling cheerful was closely linked to the commonly expressed idea of looking on the brightside, thinking positively and carrying on.

DiF results for the WEMWBS found that older adults were more likely to respond lower (signalling worse wellbeing) to feeling optimistic about the future and having energy to spare than younger adults with the same underlying wellbeing. These findings are supported by the concerns raised regarding how relevant and appropriate these items are to older adults in the content validation study, outlined above. In the IRT analysis older adults were more likely to respond higher to feeling relaxed than a younger adult with the same underlying level of wellbeing. This could be explained by the fact that some participants in the qualitative study noted that life was much less stressful, and they were much more relaxed now that they were no longer working. The findings from the IRT study that older adults at the lower end of the wellbeing scale were more likely to respond lower to feeling useful and feeling interested in other adults, while older adults at the high end of the wellbeing scale were more likely to respond higher to these items, were also mirrored by issues of relevance and acceptability in the findings of the content validation study. The frailest participants, who struggled to get out and about and were therefore quite isolated, felt that these items were no longer relevant to them. They could no longer be useful to others or get out to see friends and family or their contacts had mostly passed and so they no longer felt useful or particularly interested in other people. However, the healthier more active respondents were proud to feel useful and felt that social contact and being interested in other people was central to their QoL. Being interested in new things was often interpreted in relation to technology in the qualitative study. The result that those at the lower end of the wellbeing scale were more likely to respond lower to this item is mirrored by the fact that the frailest participants reported having no interest in new things, mostly in relation to technology.

ONS-4

In the IRT analysis, happiness provided most information about the wellbeing of older adults in the ONS-4, while anxiety provided the least. These findings reflect the importance of positive thinking and looking on the brightside, which was expressed by respondents in the qualitative study, and the stoic attitude of carrying on with life

and not dwelling on negative events that cannot be controlled, which was so often expressed in relation to negatively worded emotional and mental health items, such as anxiety. The fact that the anxiety item was found to provide the lowest level of information about underlying wellbeing again links to questions raised regarding its relevance to the wellbeing of older adults in the content validation study.

DiF results for the ONS-4 found that older adults with the same level of underlying wellbeing as younger adults were more likely to respond higher to happiness and anxiety. The finding that older adults are more likely to respond higher to the anxiety question matches results from the EQ-5D-5L and SF-12v2, and again mirrors the stoic attitude of respondents in the content validity study in relation to negatively phrased mental health questions. The DiF in relation to happiness echoes the importance of looking on the brightside commonly expressed by participants.

Another issue with the anxiety question, seen in both studies, was the use of the response scale. Participants in the qualitative study often failed to notice that the end of the scale that indicated the most positive response, was reversed for this item, as it was the only negatively worded item in the scale. Therefore, participants often gave invalid responses, stating verbally that they were not anxious, but selecting a numbered response towards the end of the scale labelled “completely” anxious. This may also contribute to the lack of expected pattern seen in the ICCs of this item in the IRT study.

The way older adults think about QoL

Several key findings from the qualitative content validation stage of this thesis regarded how the way in which older adults view their health and QoL affects the way they respond to items on PROMs. The first thing to note was that older adults' expectations of their health had declined as they aged. They no longer expected to be as fit and healthy and to be able to do all of the activities they had done as younger adults. When assessing their overall health, participants also often set aside specific health issues they were experiencing, instead making the assessment in terms of overall how healthy a person they felt themselves to be. In this way, despite significant health issues, they could continue to view their health fairly positively. Another phenomenon that was seen during the interviews in this thesis was that older adults judged their health and QoL relative to those around them of a similar age and situation who were worse off. When judging their overall health or QoL, or mentioning

limitations that they experienced, participants were often quick to compare their own state to someone of a similar age who was in a much worse state and judge themselves as lucky to be in better state. All of these response behaviours signalled response recalibration. Older adults' benchmark for what constituted good health had shifted downwards, resulting in them often responding more positively than would be expected.

It was noted in the discussion of the qualitative findings that this resulted in a risk that, in comparison with a younger person in the same health state, the older respondent would rate that same state much higher. This pattern in responding may account for many of the findings of the DiF analysis in this thesis, which found that older adults with the same underlying health as younger adults were expected to score higher on many of the questions about their health on the EQ-5D-5L and SF-12v2. This DiF resulted in higher overall expected scores for older adults with the same underlying health as younger adults on the two health measures tested.

Additional domains suggested to improve comprehensiveness

Participants suggested additional dimensions, which could improve the comprehensiveness of each and any of the instruments for older adults. These covered aspects such as: relationships, social contact and loneliness; coping and support; security about the future; control over daily life; and the way people were treated. These suggestions compare well with dimensions of the ASCOT which were found to be most closely related to the SCRQoL of older adults in the IRT analysis. The ASCOT social participation item corresponds to the suggested concepts of relationships, social contact and loneliness; the control and occupation items align with the concepts of control over daily life; and the personal and accommodation comfort and cleanliness items correspond with the suggestion of an item about coping. It is therefore likely that this measure would be found to well reflect the broader elements of QoL which are important to the QoL of older adults and it may therefore perform well in terms of content validity.

This integrated results section shows that the findings of the two separate studies within this thesis triangulate well. Similar findings were seen across the two studies, which helps to both increase confidence in the results of each of the studies individually, as well as explain potential causes of the response behaviour identified.

This reinforces the idea of the importance of using qualitative cognitive interviewing techniques to delve deeper into psychometric issues identified using statistical techniques, in order to gain an understanding of the causes of issues and therefore find the best way to solve them.

6.3 Contributions to existing knowledge

The findings of this thesis and the studies within it offer important contributions to the existing literature and knowledge regarding the psychometric performance of the included measures in older adults, the way older adults think about their QoL and respond to items on PROMs and the way in which QoL should be measured in older adults. The contributions of each study are outlined in turn below, followed by a discussion of the broader contribution of this work.

6.3.1 Systematic review

The systematic review in Chapter 3 presents a comprehensive overview of the existing evidence on the psychometric performance of the EQ-5D, SF-12, ASCOT, WEMWBS and ONS-4 in measuring the health, QoL and wellbeing of older adults. This review contributes to knowledge by identifying, appraising and synthesising the existing evidence on the performance of key measures of health, QoL and wellbeing in older adults. This population is important as they are often underrepresented in outcomes research, despite their disproportionately high use of health and social care resources.

The findings of this review provide important information on both the psychometric performance of the chosen measures in older adults and the aspects of psychometric performance for which more evidence is required. Strong evidence of the construct validity, in terms of known group and convergent validity, of the EQ-5D-3L, SF-12 and ASCOT in older adults was identified. However, there was a lack of studies investigating DiF, another important element of construct validity, with only one study found which explored age related DiF, for the SF-12 in the USA (Fleishman and Lawrence, 2003). The only examination of content validity in older adults was for the EQ-5D-3L and ASCOT. However, this was based on one study conducted on the Dutch versions of these measures in the Netherlands (van Leeuwen, Jansen et al.,

2015) and therefore with translation changes and differences in cultural attitudes towards health and QoL, these results may not be generalizable to the UK setting (Fayers and Machin, 2016), leaving no evidence of the content validity of the English versions of any of the measures in older adults. The evidence for other psychometric properties was mostly either limited, conflicting or none was identified, meaning we cannot be sure from the existing evidence about the performance of these measures in older adults in terms of structural validity, internal consistency, test-retest reliability, inter-rater reliability and responsiveness. Therefore, more evidence on the performance of the EQ-5D, SF-12 and ASCOT in older adults in relation to any of the psychometric properties, except perhaps convergent and known-group validity, was going to be of value.

An important gap in the existing evidence base is evident from the findings of this review, as no studies were found to assess the psychometric properties of either the WEMWBS or the ONS-4 specifically in older adults. This is particularly important, as interest in the use of wellbeing measures is increasing in economic evaluation. The NICE guidance on the economic evaluation of social care interventions state that wellbeing measures may be appropriate for assessing the benefits of such interventions on service users (National Institute for Health and Care Excellence, 2016). A large proportion of social care users are aged 65+. However, this review demonstrates that there is a lack of evidence on the psychometric measurement performance of such measures in older adults. It is important that wellbeing measures such as the WEMWBS and ONS-4, which are currently being used in evaluations, are suitable in the population in which they are being used. More evidence was certainly needed on their psychometric performance in older adults.

While some important conclusions can be drawn about the performance of these measures in older adults in relation to some aspects of psychometric performance, the review also identified some key weaknesses in the existing evidence in terms of both the quantity and quality of evidence available. As discussed, studies have often failed to move beyond assessments of known group and convergent validity to other crucial aspects of psychometric performance such as assessment of DiF, content validity, internal consistency, reliability and responsiveness. Studies have also focussed on CTT methods and largely ignored the benefits and additional analysis that can be conducted using IRT methods. This review contributed to knowledge by both summarising the existing evidence and outlining where this thesis should focus

to maximise its contribution to the literature on the psychometric performance of key measures of health, QoL and wellbeing in older adults.

6.3.2 Investigation of psychometric performance using item response theory methods

This study has value both methodologically and empirically. Its methodological value over much of the existing evidence identified in the literature stems from the use of IRT methods to examine psychometric performance, which offer important advantages over the more widely adopted CTT methods. Empirically it adds important information to the existing evidence base regarding the performance of the chosen measures in older adults.

As seen in the systematic review, studies to date have largely focused on CTT methods for assessing psychometric properties, with the exception of structural validity which is commonly assessed using factor analytic methods and one study which used structural equation modelling to assess age related DiF (Fleishman and Lawrence, 2003). IRT methods offer some important alternative insights and advantages over CTT methods when assessing psychometric performance. The first advantage of IRT methods regards the measurement of internal reliability and precision of measurement. CTT methods assume that internal reliability and the standard error of measurement around patients' scores are constant, regardless of the individuals' amount of the latent trait, but precision of measurement is known to vary by trait level (Hays, Staquet et al., 1998). IRT provides estimates of internal consistency and standard error of measurement which vary by trait level, enabling the researcher to understand over what range of underlying health, QoL or wellbeing the measure provides a precise measurement and where the measure may need to be improved. The second advantage of IRT methods is that they enable a more detailed investigation of the performance of item response levels and how they are used by respondents. In CTT this can only be investigated through response distributions, however in IRT, ICCs allow the researcher to clearly see the probability that each response option will be chosen by respondents at each point on the underlying trait. This allows the researcher to investigate whether there are issues in the way response categories are used such as focussing effects, misunderstanding of level labels, or levels which are indistinct from neighbouring categories. Finally, IRT methods also allow for the assessment of DiF. The presence of DiF means that the property of

measurement invariance, an important aspect of construct validity does not hold. The presence of DiF indicates that peoples' scores are not solely determined by their level of the trait but are also dependent on their demographic characteristics. This may cause bias in scores and in any resulting decisions based on those scores. It is therefore important to measure and account for any DiF. These advantages that IRT methods provide over CTT studies mean that this thesis study was able to further contribute to knowledge by providing important additional pieces of evidence on the performance of these measures in older adults, which were largely missing in the existing evidence base.

This IRT study is the first study to examine the psychometric performance of the WEMWBS and ONS-4 specifically in older adults. This evidence is vital in the current policy context in the UK. The WEMWBS and ONS-4 are two of the most widely used wellbeing measures in the UK, both of which have been included in various large surveys and local evaluations. As previously discussed, wellbeing measures are mentioned as potentially appropriate outcome measures in the evaluation of UK social care services in the NICE social care economic evaluation guidelines (National Institute for Health and Care Excellence, 2016). Older adults make up approximately half of social care spending in England and therefore we need to be sure that outcome measures used in the evaluation of social care appropriately, validly and comprehensively reflect the outcomes that are important to older adults.

The findings of this study provide important information on the psychometric performance of the selected measures. The performance of measures on different aspects of measurement performance varied and issues were identified for all measures, to varying degrees. The WEMWBS, ONS-4 and SF-12v2 were found to be internally consistent and to discriminate well across a broad range of respondents. However, the EQ-5D-5L and ASCOT displayed substantial ceiling effects for above average respondents, which resulted in reduced internal reliability and ability to discriminate the QoL of these respondents. This may cause issues when assessing interventions aimed at individuals with a low burden of disease, where expected utility values are high, as incremental effectiveness estimated using these measures will be underestimated if a proportion of individuals already rate themselves at the ceiling of the measure.

There was strong suggestion of item redundancy within the SF-12v2 multi-item scales and possible suggestion of redundancy in the WEMWBS. There were

occasional issues with the use of some response options across all the measures, but issues were more widespread in certain measures. In the ONS-4 the eleven response options available did not appear to be used evenly. Respondents were drawn to either end of the scale and five in the centre. This suggests that there are simply too many options to choose from and that they may not be being used as a smooth scale as intended.

A particularly important finding was the presence of substantial DiF in the SF-12v2 and EQ-5D-5L. The impact of this DiF was particularly strong in respondents with below average QoL. This will create issues when using results from these measures to evaluate interventions and make resource allocation decisions. The possible impact of DiF on economic evaluation analyses based on the EQ-5D-5L was investigated by exploring differences in the estimates of effectiveness and incremental effectiveness that would be received by different age groups who in fact received the same underlying gain in health in a series of hypothetical trials. This analysis revealed some important findings which contribute to the understanding of the impact of DiF on the results of economic evaluations. DiF resulted in differences in the estimate of effectiveness and incremental effectiveness received by members of different age groups who in fact received the same underlying health gain. The direction and size of these differences depended on the individuals' position on the underlying trait. Bias in the effectiveness estimates of different treatment groups, within the same age group did not cancel to lead to equal incremental effectiveness estimates as may have been expected. Therefore, DiF leads to differing ICERs according to age group membership which will bias decision making.

Bias in scores of different age groups could affect decision making in many different ways. Within an evaluation for an intervention aimed at patients with a broad range of ages, it could cause different age groups to receive inappropriately different estimates of effectiveness. If subgroup analysis is conducted, this could result in the intervention being inappropriately denied to individuals in which it is actually cost-effective, based on their age. Conversely DiF could also lead to the overestimation of effectiveness and incremental effectiveness, leading to it being inappropriately provided and resources being inefficiently used. If the intervention is only aimed at a single age group, the effectiveness estimates could simply be lower or higher than they should in fact be, potentially leading to similar errors in decision making. At the NHS level interventions, which may only be appropriate for different age groups compete for

funding. Therefore, bias in effectiveness estimates could unfairly bias funding decisions for or against certain age groups.

6.3.3 Qualitative investigation of content validity

Content validity is of critical importance to the validity of survey data, which depends upon a shared understanding between developers and respondents of the items and response options (Mallinson, 2002). The systematic review revealed a lack of content validation of the chosen measures in older adults from a UK perspective. This study contributes to knowledge by investigating the content validity of the UK versions of the EQ-5D-5L, SF-12v2, WEMWBS and ONS-4 in older adults. Cognitive interviewing techniques of think-aloud and verbal probing were used as these methods have been argued to provide the most information about the response processes which participants go through when answering survey questions. These methods are therefore best placed to identify issues which may threaten the validity of data obtained from the included measures.

In order to understand how best to comprehensively measure the QoL of older adults it is important first to understand how they conceptualise QoL and what is important to their QoL. Factors that were important to their QoL centred around their health, their ability to carry out their usual activities, social contact and emotional wellbeing. Health was essential to their QoL as it determined their ability to do their usual activities, get out and about and do what they wanted. Their ability to carry out their usual activities, both in and outside of the home was important to their QoL. Being able to do their regular daily activities within the home was key to their independence, while their ability to get out and about and engage in activities outside of the home was an important source of pleasure and facilitated social contact. Social contact with family, friends or people out in the community was of clear importance to all respondents and loneliness was often discussed as a big problem amongst older adults. Emotional strategies of stoicism and maintaining a positive outlook were adopted by many participants to combat the frustrations if declining health and ability to undertake usual activities independently.

This study is the first study identified to examine the content validity of the UK versions of the chosen measures using cognitive interviewing techniques in older adults specifically. Issues were found for all measures which could threaten the validity of

data obtained from these PROMs, however these issues were more widespread in some measures than others. Issues with interpretation were found for some items from all measures, with respondents commonly narrowly interpreting questions about usual activities to mean solely housework. Some of the more subjective questions from the two wellbeing measures were also misinterpreted by participants. For example, the worthwhile question on the ONS-4 was commonly interpreted in different ways, to mean either worthwhile to themselves or to society, whereas on the WEMWBS being interested in other people was commonly interpreted as being nosy. Differing interpretations have substantial impact on responses and call into question the validity of comparing responses between individuals and groups. Participants most frequently questioned the relevance to their QoL of negatively phrased mental health and emotional items on the EQ-5D, SF-12v2 and ONS-4 and questions about feeling optimistic about the future and loved on the WEMWBS. The SF-12v2 and ONS-4 layouts and questions caused some confusion.

The findings of this study highlight the way in which response shift impacts the way in which older adults assessed their health and QoL and the way they responded to items on the PROMs. Response calibration was commonly seen as older adults lowered the benchmark for what constituted good health, QoL or wellbeing by having generally lower expectations and assessing themselves relative to others of a similar age who were worse off. Response reprioritisation was also seen amongst the frailest respondents, who tended to feel that some items were no longer relevant to their QoL, as they could no longer perform many of the activities and roles associated with the items and had limited access to close social contact. These elements of response shift will impact scores and limit comparability between scores obtained from older and younger respondents. The response issues identified pose a threat to the validity of scores obtained from these measures and the impact this could have on decisions based on those scores. This study therefore represents an important contribution to knowledge.

6.3.4 Thesis as a whole

An important contribution of this thesis study is that it emphasises the importance of using a combination of qualitative and quantitative methods to assess the psychometric performance of measures of QoL. It also highlights the significance of triangulating the results obtained from these different methods, as together they can

provide greater insight into the performance of a measure than when considered separately. One notable finding from this study is that solely relying on statistical tests of psychometric performance, without consideration of content validity, can lead to entirely different conclusions regarding the performance of measures. For example, the WEMWBS was found to perform well in Chapter 4, as it was internally consistent over the broadest range of the underlying trait, exhibited minimal DiF and did not exhibit large floor and ceiling effects. However, issues with content validity in an older sample were widespread within this measure in the study presented in Chapter 5, which led to the opinion that this measure does not well reflect the wellbeing of older adults. This reinforces the importance of testing content validity in the specific population in which the measure will be used, as issues with content validity can substantially impact on the validity and quality of data obtained and any decisions based on that data.

Another important finding from the triangulation of quantitative and qualitative results in this thesis is that cognitive interviewing content validation techniques are able to provide valuable information on the likely causes of DiF identified in quantitative studies. Information on causes of DiF can enable measure developers to amend items which exhibit DiF and can provide information on the types of questions to avoid where patterns are seen across items and measures. This information is therefore of value, not only for improving current measures, but also in the development of future measures. Few studies go beyond assessments of psychometric performance to qualitative research to understand those issues and therefore this represents an important contribution to knowledge.

6.4 Implications of thesis findings

This thesis has implications for a range of stakeholders involved in both conducting and assessing economic evaluations of health and social care interventions, as well as for researchers looking to develop measures suitable for assessing the impact of health and social care interventions on the QoL of older adults. In this section of the discussion, these implications will be outlined and recommendations will be made to relevant stakeholders.

6.4.1 Implications and recommendations for economic evaluation of health and social care interventions

Differential Item Functioning

The identification of DiF in the health measures has important implications for economic evaluation, particularly given NICE's requirement for the use of the EQ-5D in the evaluation of healthcare interventions. The analysis of the impact of DiF in the EQ-5D-5L on the estimates of effectiveness and incremental effectiveness generated from hypothetical trials, in section 4.4.4, shows that this DiF does go on to bias both effectiveness and incremental effectiveness estimates and will therefore impact decision making. These findings reinforce the importance of controlling for DiF in economic evaluation.

It is therefore important that those parties involved in conducting economic evaluation control for DiF in their economic evaluation. It is also necessary that the Evidence Review Groups involved in assessing the quality of economic evaluations submitted to NICE make it standard practice to check that DiF has been controlled for in NICE submissions, or at the very least that the likely impact of DiF on the results has been considered. If this is incorporated into standard practice the NICE committee have all possible information available, with which to judge the level of bias that is likely to be present in estimates of effectiveness and incremental effectiveness as a result of DiF, which increases confidence in estimates provided and decisions based on these estimates.

Measure Choice recommendations

Recommendations surrounding which existing measure is best to use when evaluating the impact of a health and social care intervention on the QoL of older adults are of great value to companies conducting clinical trials with a view to submitting evidence of the cost-effectiveness of their treatment to NICE. It is also useful for those assessing the quality of NICE submissions to understand the advantages and disadvantages of different measures in this area, in order for them to be able to make an evidence based assessment of the methods which have been used to measure and value QoL, as this is a central component of the cost-effectiveness of a new treatment. It is also important that NICE themselves are aware of the strengths and weaknesses of the measures available to assess the impact of health and social care interventions on older adults, as there is currently no clear

guidance for companies on which measure should be used in this area. Lack of clear guidance leads to a lack of comparability between evaluations of health and social care interventions, which makes resource allocation decisions across evaluations difficult.

A single clear recommendation of which measure or combination of measures should be used to evaluate health and social care interventions in older adults is difficult to make, as various issues of validity, internal consistency and acceptability were found for each of the measures examined within this PhD research programme. However, issues identified were more widespread and posed a bigger threat to the validity of some measures than others.

Content validity is argued to be the most important element of validity (COSMIN Group, 2018), as the validity of the data received from questionnaires is dependent on whether the questions and response options are understood by the respondent, whether those questions are relevant to the concept that the instrument aims to measure and whether the important aspects of that concept are comprehensively captured by the instrument. The layout and style of question caused confusion on the SF-12v2 and ONS-4, which impacted the validity of responses received. The relevance of at least one item was questioned by some respondents on every measure, however this issue was more widespread on the WEMWBS, ONS-4 and SF-12v2 than it was on the EQ-5D-5L. The more subjective wellbeing items and negatively worded mental health items were more commonly felt to be irrelevant to older participants as they reported that they did not often think about their life in this way. This was particularly true for frailer older adults, who felt that very few of the subjective items of the WEMWBS and ONS-4 were relevant to their life anymore, as their functional abilities had declined and therefore basic functionings were more their focus, rather than broader elements of QoL connected to having a role and purpose and social connection. It was felt that participants often found the more concise and practical functioning focussed EQ-5D items easier to answer and relevant to their daily life, with the exception of anxiety/depression.

The findings from the content validity study would therefore suggest that, of the four measures tested, the EQ-5D-5L may be the best starting point for measuring the effectiveness of health and social care interventions for older adults. However, this does not mean that the EQ-5D-5L performed perfectly and several issues, beyond the need to control for DiF, need to be considered. Firstly, the EQ-5D-5L was found

to exhibit substantial ceiling effects for respondents with above average health in both age groups, resulting in reduced internal reliability of the measure in these individuals. Ceiling effects reduce the ability of the measure to precisely discriminate the exact level of QoL in these respondents. Therefore, when assessing interventions aimed at individuals with a low burden of disease or services intended as early interventions, where expected utility values may be quite high, the current EQ-5D may underestimate incremental effectiveness if a proportion of individuals already rate themselves at the ceiling of the measure.

Secondly, in the cognitive interviewing study some participants did not feel that this measure comprehensively covered what was important to their QoL. If important elements of QoL are missed, services which improve these elements may be undervalued. There are several options going forward to improve the measurement of the effectiveness of health and social care interventions for older adults:

1. Use the EQ-5D-5L in combination with another measure which covers the broader aspects of QoL that are important to older adults
2. Adapt the current EQ-5D-5L by adding bolt-on dimensions
3. Develop a new measure
4. Move towards adaptive descriptive systems

One of the suggestions in the NICE social care guidelines is to conduct a primary analysis of effectiveness based on the EQ-5D but also conduct a parallel analysis based on a broader measure of QoL, such as the ASCOT, ICECAP or a wellbeing measure. Using a combination of the EQ-5D and either the ICECAP-O or the ASCOT may be a good option in terms of coverage of concepts that are important to the QoL of older adults. As discussed in section 5.5.1, the ICECAP-O and ASCOT link well with the additional concepts suggested by participants to improve the coverage of the measures included in the content validity study. The use of a combination of measures could also help with ceiling effect issues if the other measure includes questions which require higher level functioning than the EQ-5D. However, the use of separate measures becomes problematic when it comes to converting scores from two measures into a single preference-based utility value suitable for the calculation of QALYs. The use of different measures across different evaluations or in different populations also has issues for cross-programme comparability, as different measures contain different items and therefore assess different outcomes. It would

be preferable to use a single measure in order to maintain comparability across evaluations.

The second option is to adapt the EQ-5D-5L by extending its descriptive system so that this measure better reflects what is important to the QoL of older adults. This could be done by adding bolt-on dimensions or items that have been identified as relevant to the QoL of older adults. Again, adding items which require higher level functioning in participants can also help to reduce the ceiling effect issue. While amending the descriptive system of the EQ-5D will have an impact on comparability of effectiveness estimates between evaluations, it has been argued to have less impact than using different measures altogether as the core set of items remains consistent between studies (Finch, Brazier et al., 2017). The impact of changes to the descriptive system on the value set of the measure would have to be investigated and it is likely that the preference weightings of the measure would have to be recalculated.

The third option is to develop a new measure. The advantage of this strategy is that researchers could start from scratch, with the aim of measuring broader QoL, and ground the generation of domains and items in what is important to a broad range of health and social care users, using qualitative methods to take the views of the population directly into account when generating items. These methods have recently been used to develop measures such as the ICECAP-O (Grewal, Lewis et al., 2006) and ReQoL (Keetharuth, Brazier et al., 2018) and have been argued to produce measures with superior content validity than more traditional measure development techniques, which focussed more on using the literature and expert opinion of developers, as was done for the development of the EQ-5D. Another benefit of developing a new measure is that questions requiring a range of different levels of ability and functioning can be incorporate to solve the issue of ceiling effects. During development, the presence of ceiling effects can also be investigated and resolved through amendments to the measure.

Work to develop a preference-based measure for a broader QALY, appropriate for the economic evaluation of health and social care is currently being undertaken in ScHARR (School of Health and Related Research, 2018). This study, called the eQALY project, has in fact focussed its domain and item generation in the opinion of a broad range of health and social care users and carers and has used cognitive interviewing methods to test potential items in those same groups. It is therefore

hoped that this will lead to a broad measure of QoL, which comprehensively, validly, responsively and reliably measures the impact of the range of health and social care interventions in those populations which receive them. However, the limitation of developing a new measure is a lack of comparability between assessments that have used different measures to date.

The fourth option would be to move away from standardised descriptive systems, in which everybody is asked the same questions, towards adaptive descriptive systems. Recent developments in outcome measurement have used IRT methods to develop computer adaptive tests (CATs) (Fayers and Machin, 2016). CATs are developed by calibrating a large bank of items related to a concept onto a single latent trait. Since all items are positioned on the same latent trait scale, consistent trait scores can be estimated, regardless of which questions a respondent answers (Fries, Bruce et al., 2005). The CAT algorithm uses the respondent's answer to an item to evaluate the respondent's most likely position on the trait scale. It then chooses the most informative question about that area of the trait to try and increase the precision of its estimate of the respondent's level of the trait. This process continues until a cut-off is reached for sufficient precision of measurement (Ware, Bjorner et al., 2000). These methods have a number of important potential advantages over standardised descriptive systems.

Firstly, the same precision of measurement can be obtained from fewer questions than in a standardised test (Fries, Bruce et al., 2005). This is because we can avoid asking questions which ask about areas of the trait which are likely to be irrelevant, given the answers already obtained. For example, if a respondent states that they are limited in climbing a flight of stairs, asking if they are limited in running 10km is not likely to give us much more information, as it is very likely that they will also be limited in this activity. However, going on to ask about more basic activities of daily living may give a more precise estimate of their relative position on the scale. It has been shown that the same precision of measurement can be obtained from CATs which are 30-50% shorter than standardised descriptive systems (Fayers and Machin, 2016). This would reduce respondent burden and improve data quality, as it is less likely that respondents will disengage during a questionnaire. Secondly, the items asked are likely to be more relevant to respondents, again decreasing the likelihood of respondent disengagement, invalid responses and high rates of missing data. Thirdly, by having a bank of possible items which cover a broad range of underlying trait levels, the likelihood of issues with ceiling effects are greatly reduced. Lastly, different

questions can be asked to different groups or the same question can be calibrated differently in different groups, which can avoid or account for DiF. If a question is known to be interpreted differently or to have differing importance between groups, different parameters could be used to calibrate that item and estimate the trait scores of members of different groups. If a question is known to be important in a certain group but irrelevant in another, it could only be asked in the relevant group. This reduces the likelihood that DiF will bias scores and any resulting decisions based on these and increases the comprehensiveness of measures for groups with different priorities or conceptualisations of QoL.

These advantages of adaptive testing would solve some many of the issues identified in this thesis. As outlined, issues of age related DiF, which could go on to causing bias in estimates of effectiveness of interventions, could be avoided. Secondly, issues with the relevance of items for older adults, which continue to get worse as those respondents experience more severe frailty could be avoided by selecting items from relevant areas of trait and items known to be important to the QoL of older adults. Lastly, issues with the lack of coverage of important aspects of QoL which health and social care services for older adults may seek to improve, which if missed may result in these services being undervalued and underfunded could be solved as these items could be asked to older adults, but not younger adults.

Whichever measure or method chosen, it is important to conduct a thorough evaluation of the presence and impact of DiF to ensure that any estimates of effectiveness and funding decisions based on an economic evaluation using the chosen measure are not biased towards or against groups with certain characteristics. It is also important to ensure that the aspects of QoL which are important to the target population of the measure and which may be impacted by services they may receive are comprehensively covered by any PROM taken forward.

6.4.2 Implications and recommendations for measure development

The results of this thesis provide some important recommendations for the measurement of the QoL of older adults, whether this measurement takes the form of a standardised or adaptive descriptive system. These recommendations are of value to anyone involved in the development of measures or adaptive tests of QoL suitable

for assessing the impact of health and social care interventions on the QoL of older adults.

Firstly, this study identified key aspects of QoL which are important to the QoL of older adults and should be included in assessments of the effectiveness of health and social care interventions aimed at older adults. Central aspects of QoL were found to be people's health, ability to carry out their usual activities and social contact and emotional wellbeing. Health was viewed as a mechanism through which they were able to carry out activities that they valued and enjoyed, get out and about and make regular social contact with family, friends and members of the community. The importance of regular social contact was clear as isolation and loneliness were often discussed as one of the biggest issues facing older adults. Other elements of QoL which were important to older adults were independence, control over daily life, support/coping and security, both financial and surrounding the availability of support. As health is viewed as important by older adults, not in itself but through its ability to enable them to achieve broader elements of QoL such as social participation, independence and control over their daily life, it is these elements of QoL that health and social care services aimed at older adults aim to maintain or improve. It is therefore crucial that these broader elements of QoL are included in any measure of QoL being used to measure the effectiveness of such interventions or else they may be undervalued and underfunded.

Secondly, this thesis identified some patterns in the way that older adults responded to the measures, which could be of use when generating items for a new measure. The phrasing of certain concepts should be considered carefully. The stoic attitude towards issues among the older population means that they are reluctant to signal issues to negatively worded mental health items. This reaction seemed particularly strong for items which specifically referenced anxiety or depression. Participants seemed more comfortable with terms such as feeling down, which may be more socially acceptable to them. Therefore, future measure developers may either want to focus on positively worded mental health items or to phrase negatively worded items carefully, avoiding more official or severe terms such as anxiety and depression. Another concept which needs to be phrased carefully is energy. Items aimed at older adults should be phrased in terms of whether they feel they have enough energy, rather than plenty of energy as this was felt to be unrealistic by many participants. If developers wish to ask about respondents' ability performing activities, they should

be sure that if they provide specific examples, these are relevant to the ability level of a broad range of older respondents.

Measure developers should also avoid ambiguous and overly subjective items where possible as this leads to interpretation issues. Older respondents in this study tended to prefer measures with clear and succinct items and items with more of a functional focus. Therefore, focussing on this type of phrasing may improve respondent engagement and the validity of data received. Finally, measure developers should be aware of response shift, particularly in global assessments of general health, where respondents often interpreted best health imaginable or excellent health to be in relation to what they would expect for someone of their age. Measure developers may want to consider additional instruction on the interpretation of anchors to eliminate or reduce this effect.

These recommendations could be of great value for the development of items for a new QoL measure for use in older adults or for the selection of existing items which may perform well or be appropriate in older adults for use in an adaptive system.

6.5 Limitations

This thesis and the studies within it have some important limitations, which need to be discussed.

The first thing of note is that it was not feasible within this PhD study to investigate all aspects of psychometric performance in each of the measures. This study certainly does not claim that the only important aspects are those examined within this thesis. Other important elements such as responsiveness and test-retest reliability are also important to test to ensure a well performing measure however these elements require experimental longitudinal data. While some of the datasets used in the first phase of this PhD come from longitudinal surveys, the datasets cannot be used as panel datasets which link data from the same individuals annually. However simply using panel datasets collected annually, would not be appropriate as responsiveness measures whether an instrument can detect change over time where change is known to have occurred, and we may not be able to accurately judge this using panel

datasets while test-retest requires repeated measurements over relatively short time periods, assuming no change has occurred and the time periods between panel measurements would be far too long to be able to be sure that no change had occurred. Therefore, tests of these aspects of measurement performance often require primary data collection in fairly large groups of respondents which was not feasible in this thesis study. For now, we will have to rely on existing evidence found in the systematic review for the EQ-5D and the SF-12 and note that this is an important area of future research, particularly for the WEMWBS and ONS-4 for which no evidence was found in older adults.

Similarly, this study includes only a very small selection of the available generic measures of health, QoL and wellbeing which could be appropriate for measuring these concepts in older adults, and again we do not claim that these are the only available or possibly appropriate measures. There are hundreds to choose between. This study came from a starting point of NICE's current practices in the economic evaluation of healthcare interventions and extending these to social care evaluation, using the case study of older adults as they reflect a large proportion of service users in both health and social care. Therefore, while accounting for the fact that solely measuring health may not adequately reflect all the outcomes of social care interventions, we felt it was still important to attempt to find a single generic measure, suitable in all groups, but with a broader perspective than solely health, which may adequately reflect all outcomes for both older and younger adults, therefore ensuring comparability between evaluations. While many measures aimed specifically at measuring the QoL of older adults exist such as the ICECAP-O, this would mean that different measures would be used in different age groups and comparability between interventions may not be maintained and therefore such measures were not included in this study, even though the ICECAP measures are included in the NICE social care guidelines. However, these measures may do a better job of assessing the QoL of youngers and older adults separately and not using them may sacrifice accurate and comprehensive assessment for the sake of comparability. Future work may be needed to investigate how much we are losing, in terms of the accuracy and comprehensiveness of measuring what is important to older adults, by attempting to maintain comparability through using a single measure in all groups.

Some of the limitations of this study are related to the datasets used in Chapter 4 to assess psychometric performance using IRT methods. One important limitation of large datasets such as these is that they often do not include the frailest, who are less

likely to respond and take part. There was also very limited, or no coverage, of those living in nursing and residential homes, who make up an important group of the frailest older adults in two of the datasets used. The ASCS, as a survey of social care users, included older adults living in residential and nursing homes however the HSE interviewed only community dwelling individual and only 1.4%. Therefore, the sample of older people in these datasets and in the resulting analysis may not be representative of the older population in the UK and, most notably, may be healthier than the older population in the UK.

Other limitations may relate to the choice of recruiting through the CARE 75+ cohort for the content validation study. While it was felt that recruiting through an existing and ongoing cohort would improve recruitment rates in a group who are known to be difficult to reach and underrepresented in research, particularly those older adults experiencing more severe frailty, recruiting through the cohort may have had an impact on the results obtained. The CARE 75+ study involved visiting participants in their homes repeatedly over a five-year period and asking many questions about their health and QoL. Participants were therefore somewhat accustomed to being asked questions from measures such as those included in this study. They therefore may be more accepting of such questions than older adults more generally, who are unused to PROMs. It is likely that older adults recruited to the cohort, who found questions such as these upsetting, irrelevant or inappropriate were more likely to have dropped out of the cohort and therefore their details would not have been passed on for recruitment to this study. The CARE 75+ study also did not recruit older adults living in care homes. This may miss those experiencing the most extreme frailty, whose views on measures such as these are also important to obtain as they are intensive users of health and social care services.

6.6 Future research

The findings of this thesis and its limitations offer some useful suggestions for future research.

As discussed in the limitations section of this thesis, not all aspects of the psychometric performance of the included measures in older adults could be tested in this study. Reliability and responsiveness are also important elements of psychometric performance which require testing to ensure that estimates of health,

QoL or wellbeing obtained from such measures are stable where no change in status has occurred and that they return appropriately difference scores, which reflect the change in state, when change does occur. Evidence of the reliability and responsiveness of the measures included in this study was generally found to be limited at best (with the exception of moderate evidence of the responsiveness of the EQ-5D-3L, which was mostly based on studies in older adults who had recently experienced hip and femoral neck fractures) in the systematic review. It is very important that further research is dedicated to ensuring the responsiveness of these measures in older adults in relation to a range of health and social care interventions. If a measure is to be recommended in the NICE guidelines for broadly evaluating social care and/or health interventions, it needs to be responsive in a broad range of those interventions likely to be offered or evaluated. As discussed, the aims of different interventions, even within one of these sectors can vary substantially, let alone across sectors. Therefore, it is important to find a measure which accounts for the outcomes of a broad range of services and reports appropriate change scores in order to adequately evaluate services and ensure optimal resource allocation decisions to be made. This is especially important as wellbeing measures in general populations have been found to be less sensitive to change than sector specific measures such as the EQ-5D and ASCOT (Mukuria, Rowen et al., 2016).

While a range of older adults, in terms of health and frailty, were captured in this research, it was also felt that some important groups were missed. The frailest participants in the qualitative study often had notably different reactions to items in the measures than the healthier participants. It was noted, particularly for the wellbeing measures, that the frailest participants questioned the relevance of more items to their QoL. The content validation study conducted in this thesis only included community dwelling older adults. Therefore, the opinion of older adults living in residential and nursing homes was missed. With the exception of the ASCOT in the IRT study, older adults in non-community settings were also largely missed. However, these older adults represent the frailest members of the older population and are intense users of health and social services. Therefore, their opinion and the performance of measures in this group is important. Future research should investigate the psychometric performance of existing measures of health, QoL and wellbeing in older adults living in nursing and residential homes.

An important focus of future research surrounds what concepts a broader QALY needs to include. Whilst broadly all the measures included in this research seek to

measure QoL in some form, they differ in terms of their focus and the specific concepts included, with EQ-5D and SF-12 focussing on health, ASCOT on the impact of social care on QoL, WEMWBS on mental wellbeing and ONS-4 on personal subjective wellbeing. These measures cannot be directly compared in terms of performance without additional qualitative consideration of what should be included in a broader QALY. It is important that the content and focus of this broader QALY aligns with the policy and service perspective which it is being used to evaluate, otherwise the impacts of these services will be missed, and they will continue to risk being undervalued and underfunded. While there are regular arguments for broadening the QALY beyond health, further work needs to be carried out to decide exactly what concepts are important to include in a comprehensive assessment of broader QoL and wellbeing and the full breadth of services which this broader QALY will be used to evaluate needs to be considered to be sure that the resulting measure is appropriate.

This broader QALY could be based on an existing measure, which may or may not require adaptation in terms of content, or a new measure could be developed. The choice of potentially appropriate existing measures on which to base future economic evaluation using a broader QALY is not limited to those tested in this study. The psychometric performance of other measures not tested, such as the ICECAP measures could be evaluated both quantitatively and qualitatively. The content validity of other available measures such as ICECAP and ASCOT could also be assessed as the content of these broader measures closely align with additional areas which were suggested by respondents to be important to improve the comprehensiveness of the health and wellbeing measures tested in this research. An interesting piece of future research would be to repeat the qualitative content validation study conducted in this thesis using the EQ-5D-5L, ICECAP-O and ASCOT in order to further investigate response issues experienced in these measures as well as respondents' preferences for measures and how well they think each measure is able to comprehensively reflect their QoL. In this way we could start to build a more comprehensive picture of the aspects of QoL, and the ways of asking about those aspects, which are most relevant and effective in a broader QALY measure. This way, even if a new measure has to be generated, this new measure will be based on informed research. Work to develop a preference-based measure for a broader QALY, appropriate for the economic evaluation of health and social care is now being undertaken in ScHARR (School of Health and Related Research, 2018). A broad range of health and social care users with different conditions and carers are being involved in the development of this

measure which should result in a measure with good content validity. Future research examining the performance of this measure, once it is completed, would be of great interest. Future research may also focus on the continued development of adaptive measures and how this could be applied to the evaluation of health and social care services aimed at older adults.

An important consideration in the potential use of measures in the economic evaluation of health and social care is that they need to be preference-based, and that this needs to be on an appropriate scale for any broader QALY that results. Currently, the EQ-5D is preference-based on anchors of best and worst health imaginable using time trade off (TTO) exercises in the general population (Dolan, 1997), while ASCOT is preference-based on anchors of all to none of an individual's social care needs being met using best worst scaling exercises in social care users, anchored to death by a TTO exercise (Netten, Burge et al., 2012). The WEMWBS and ONS-4 are not currently preference-based, while the developers of the SF-12v2 state that it is preference-based using IRT methods (Maruish, 2012). Any future decision broadening the QALY may therefore involve not only a change of measure to one which comprehensively captures those aspects of QoL and wellbeing which have been found to be important to the broader QALY, but also an accompanying preference elicitation using appropriate methods, in an appropriate sample, using appropriate anchors for the resulting broader QALY. The current eQALY work in ScHARR will include the generation of a value set in order to make the final measure preference-based (School of Health and Related Research, 2018).

While it may not always be possible to construct a measure which is free from DiF in all its items and in relation to all respondent characteristics, it is important that future research investigates methods for controlling for DiF, for example through the use of MIMIC modelling (Fleishman and Lawrence, 2003) or anchoring techniques (Knott, Lorgelly et al., 2017). If methods for controlling for DiF could become part of the outcome measurement process in economic evaluation this would reduce or remove this source of bias in effectiveness estimates and decision makers could proceed in making resource allocation decisions without this additional concern.

6.7 Conclusion

This thesis provides some important information on the psychometric performance of a selection of health, QoL and wellbeing in older people.

This thesis found that there were large and important gaps in the existing evidence on the psychometric performance of existing and commonly used measures of health, QoL and wellbeing in older adults. Studies to date have often focussed on CTT tests of construct validity in terms of convergent and known group validity. There was limited evidence, in terms of both the quality and quantity of available evidence, on the content, structural and construct (DiF) validity, reliability and responsiveness of many of the included measures. The vast majority of evidence focussed on CTT methods despite the advantages of IRT methods, which enable the study of DiF, detailed information on the performance of item levels and estimation of internal consistency reliability and measurement error at any point on the underlying health, QoL or wellbeing scale.

The quantitative and qualitative elements of this thesis came together to provide valuable insights into the psychometric performance of the included measures. Issues were found for all measures. The EQ-5D-5L and ASCOT exhibited substantial ceiling effects for above average respondents in both age groups, resulting in reduced internal reliability and reduced ability to precisely discriminate the QoL of these respondents. Item redundancy was noted in both studies within the SF-12v2 multi-item scales, resulting in a TLA super item approach being taken for this measure in the IRT study. The two health measures tested, the EQ-5D-5L and SF-12v2, both exhibited substantial DiF in relation to age. The likely impact of the DiF identified on decision making, through the introduction of bias into estimates of effectiveness and incremental effectiveness, on which resource allocation decisions, are based was demonstrated. This will cause bias both within individual appraisals and when funding decisions are being made between different appraisals aimed at different age groups. This finding reinforces the need for future research to focus on ways to control for DIF in routine practice within economic evaluation.

There were occasional issues with the use of some response options across all the measures. In the ONS-4 the eleven response options available did not appear to be used evenly, with respondents drawn to either end of the scale and five in the centre. This suggests that there are simply too many to choose from and that they may not be being used as a smooth scale as intended. In the content validation study, there

were also problems with inconsistent responding in the ONS-4, as participants failed to notice the reversal of the response scale for the negatively worded anxiety item, which may have contributed to the lack of expected pattern in the IRT study. In the SF-12v2, even scores dominated the ICCs of the super items, suggesting that there is a strong tendency for respondents to choose the same response option for each item within a super item pair, resulting in even scores. The layout of the SF-12v2 and length of the questions also sometimes caused confusion in the content validation study and led some participants to prefer other, more measures.

Issues of the relevance of items to older adults and the comprehensiveness of each of the measures were also widely noted in the cognitive interviewing study. While the cognitive interviews identified specific response issues for each measure; general patterns were also seen across measures which can lead to broader recommendations for researchers looking to develop measures of QoL aimed at older adults. Participants most frequently questioned the relevance to their QoL of negatively phrased mental health and emotional items as these items did not fit with the stoic generational attitude of not dwelling on issues which could not be controlled and looking on the brightside. Participants also questioned the relevance of some of the more subjective wellbeing items as they said that they did not think about their lives in this way. These participants preferred questions which focussed on their ability to function in key areas of their life.

The content validation study also provided important information about the way in which response shift impacts the way older adults respond to PROMs. Response recalibration was closely linked to the way that older adults viewed their health and QoL, as they described having lower expectations of their health than when they were younger and therefore, despite significant health issues they often continued to view their state positively. Participants also often judged their own state relative to other members of their age group who they knew were worse off, again enabling them to continue to view their own state positively. Response reprioritisation was also seen in the frailest respondents, who in response to declining physical functioning chose to focus on their ability to carry out more basic functionings, shifting their priority away from activities which required higher functionings. In this way again, they could still rate themselves fairly highly on some items. These findings emphasise the risk that, in comparison to a younger person in the same health state, older respondents will rate the same state much higher. This could be an important source of DiF across measures.

It was clear from this study that none of the included measures provided a comprehensive view of the QoL of older adults. The coverage of any of the measures would need to be extended to include broader elements of QoL identified as important to older adults, such as social contact and independence. Participant's preference for a concise, functional focussed measure suggested that the EQ-5D-5L was the preferred measure of those included. Therefore, this could be used as a starting point, from which adaptions to the EQ-5D-5L could be made, or a new measure developed based on the EQ-5D style.

This study provides important contributions to the existing knowledge on the psychometric performance of the included measures in older adults. It is therefore a useful source of information for evaluators seeking to choose an appropriate existing measure for use in the evaluation of health and social care services in older adults. The study also goes beyond this to explore the way in which older adults think about their health, QoL and wellbeing and how this affects the way they respond to PROMs. This information is of great value for researchers aiming to develop a new measure of QoL suitable for the evaluation of interventions aimed at older adults. Future research can use this information as a starting point from which to refine existing PROMs or develop new psychometrically superior measures of QoL in this population.

References

- Abanobi, O. C. (1986). Content validity in the assessment of health status. *Health Values* 10(4) 37-40.
- Adult Social Care Statistics and NHS Digital (2016). Personal Social Services: Expenditure and Unit Costs England 2015-16.
- Age UK. (2018a). *Changes to state pension age* [online]. Available at: <https://www.ageuk.org.uk/information-advice/money-legal/pensions/changes-to-state-pension-age/> [Accessed 7th November 2018].
- Age UK (2018b). Later Life in the United Kingdom.
- Al-Janabi, H., T. N. Flynn and J. Coast (2012). Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Quality Of Life Research: An International Journal Of Quality Of Life Aspects Of Treatment, Care And Rehabilitation* 21(1) 167-176.
- Alshreef, A. and S. Dixon (2015). Wellbeing Measures - Scoping document to inform the evaluation of NHS Vanguard sites.
- Anderson, C., et al. (2000). Home or hospital for stroke Rehabilitation? Results of a randomized controlled trial : II: cost minimization analysis at 6 months. *Stroke* 31(5) 1032-1037.
- Ariss, S. M., et al. (2015). Secondary analysis and literature review of community rehabilitation and intermediate care: an information resource.
- Bansback, N., et al. (2012). Using a discrete choice experiment to estimate health state utility values. *J Health Econ* 31(1) 306-318.
- Barbour, R. (2010). Focus Groups. In I. Bourgeault ed. *SAGE Handbook of Qualitative Methods in Health Research*. London, SAGE Publications.
- Beaton, D. E., et al. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* 25(24) 3186-3191.
- Bennett, L. and R. Humphries (2014). Making best use of the Better Care Fund. Spending to save? 1-16.
- Bentur, N. and Y. King (2010). The challenge of validating SF-12 for its use with community-dwelling elderly in Israel. *Quality Of Life Research: An International Journal Of Quality Of Life Aspects Of Treatment, Care And Rehabilitation* 19(1) 91-95.
- Bice, T. W. (1976). Comments on health indicators: methodological perspectives. *Int J Health Serv* 6(3) 509-520.

BioMed Central Ltd (2019). ISRCTN Registry.

Bjorner, J., M. Kosinski and J. Ware (2003). The feasibility of applying item response theory to measures of migraine impact: a re-analysis of three clinical studies. *Qual Life Res* 12(8) 887-902.

Braun, V. and V. Clarke (2006). Using thematic analysis in psychology. *Qualitative research in psychology* 3(2) 77-101.

Brazier, J., et al. (2014). 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* 18(34) vii-viii, xiii-xxv, 1-188.

Brazier, J., et al. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford: Oxford University Press.

Brazier, J. and A. Tsuchiya (2015). Improving Cross-Sector Comparisons: Going Beyond the Health-Related QALY. *Applied Health Economics And Health Policy* 13(6) 557-565.

Brazier, J. E. and J. Roberts (2004). The estimation of a preference-based measure of health from the SF-12. *Med Care* 42(9) 851-859.

Brazier, J. E., et al. (2012). Developing and testing methods for deriving preferencebased measures of health from condition-specific measures (and other patient-based measures of outcome). *Health Technology Assessment* 16(32) 1-109.

Brazier, J. E., et al. (1996). Using the SF-36 and Euroqol on an elderly population. *Qual Life Res* 5(2) 195-204.

Brod, M., L. E. Tesler and T. L. Christensen (2009). Qualitative research and content validity: developing best practices based on science and experience. *Qual Life Res* 18(9) 1263-1278.

Brooks, R., R. Rabin and F. de Charro (2003). *The measurement and valuation of health status using EQ-5D: A European perspective*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brooks, R. and G. The EuroQol (1996). EuroQol: the current state of play. *Health policy (Amsterdam, Netherlands)* 37(1) 53-72.

Buers, C., et al. (2014). The value of cognitive interviewing for optimizing a patient experience survey. *International Journal of Social Research Methodology* 17(4).

Bulamu, N. B., B. Kaambwa and J. Ratcliffe (2015). A systematic review of instruments for measuring outcomes in economic evaluation within aged care. *Health Qual Life Outcomes* 13 179.

Cappelleri, J. C., J. Jason Lundy and R. D. Hays (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther* 36(5) 648-662.

Cartwright, M., et al. (2013). Effect of telehealth on quality of life and psychological outcomes over 12 months (Whole Systems Demonstrator telehealth questionnaire study): nested study of patient reported outcomes in a pragmatic, cluster randomised controlled trial. *BMJ (Clinical Research Ed.)* 346 f653-f653.

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioural Research* 1(2) 245-276.

Cernin, P. A., et al. (2010). Reliability and validity testing of the Short-Form Health Survey in a sample of community-dwelling African American older adults. *Journal of Nursing Measurement* 18(1) 49-59.

Chang, C. H. and B. B. Reeve (2005). Item response theory and its applications to patient-reported outcomes measurement. *Eval Health Prof* 28(3) 264-282.

Clarke, A., et al. (2011). Warwick-Edinburgh Mental Well-being Scale (WEMWBS): validated for teenage school students in England and Scotland. A mixed methods assessment. *BMC Public Health* 11 487.

Clegg, A., et al. (2013). Frailty in elderly people. *Lancet* 381(9868) 752-762.

Coast, J., et al. (2008). Valuing the ICECAP capability index for older people. *Social Science & Medicine* 67(5) 874-882 879p.

Coast, J., et al. (1998). Use of the EuroQoL among elderly acute care patients. *Qual Life Res* 7(1) 1-10.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed ed. Hillsdale, NJ: Earlbaum.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Qual Life Res* 12(3) 229-238.

COSMIN Group. (2018). *Cosmin Taxonomy of Measurement Properties* [online]. Available at: <https://www.cosmin.nl/tools/cosmin-taxonomy-measurement-properties/> [Accessed 27 December 2018].

Culyer, A. (1983). Introduction. In A. Culyer ed. *Health Indicators: An International Study for the European Science Foundation*. New York, St Martins. 1-22.

Curry, N. and C. Ham (2010). Clinical and service integration The route to improved outcomes. *The Kings Fund* 1-64.

Davidson, S., et al. (2009). Well? What do you think? 2008 - The Fourth national Scottish survey of public attitudes to mental wellbeing and mental health problems. In S. G. S. Research ed.

Davis, D. H., et al. (2015). Montreal Cognitive Assessment for the diagnosis of Alzheimer's disease and other dementias. *Cochrane Database Syst Rev*(10) CD010775.

Davis, J. C., et al. (2012). Exploration of the association between quality of life, assessed by the EQ-5D and ICECAP-O, and falls risk, cognitive function and daily function, in older adults with mobility impairments. *BMC Geriatr* 12 65.

Department for Constitutional Affairs (2007). Mental Capacity Act 2005: Code of Practice. The Stationery Office, London.

Department of Health. (2013). *2010 to 2015 Government Policy: Health and Social Care Integration* [online]. Available at: <https://www.gov.uk/government/publications/2010-to-2015-government-policy-health-and-social-care-integration/2010-to-2015-government-policy-health-and-social-care-integration>.

Devlin, N. J. and R. Brooks (2017). EQ-5D and the EuroQol Group: Past, Present and Future. *Appl Health Econ Health Policy* 15(2) 127-137.

Devlin, N. J., et al. (2017). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Econ.*

Diaz-Redondo, A., et al. (2014). EQ-5D rated by proxy in institutionalized older adults with dementia: psychometric pros and cons. *Geriatr Gerontol Int* 14(2) 346-353.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Med Care* 35(11) 1095-1108.

Dolan, P., R. Layard and R. Metcalfe (2011). Measuring Subjective Well-being for Public Policy. Office for National Statistics.

Dolan, P. and R. Metcalfe (2012). Measuring Subjective Wellbeing: Recommendations on Measures for use by National Governments. *Journal of Social Policy* 409-427.

Drummond, M., et al. (2005). *Methods for the Economic Evaluation of Health Care Programmes*. 3rd ed: Oxford University Press.

Dubos, R. (1979). Introduction. In N. Cousins ed. *Anatomy of an Illness as Perceived by the Patient*. New York, Newton. 11-23.

Ebrahim, S. (1995). Clinical and public health perspectives and applications of health-related quality of life measurement. *Soc Sci Med* 41(10) 1383-1394.

Edelen, M. O. and B. B. Reeve (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res* 16 Suppl 1 5-18.

Elbers, R. G., et al. (2012). Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: a systematic review of measurement properties. *Qual Life Res* 21(6) 925-944.

Fayers, P. and D. Machin (2007). *Quality of life : the assessment, analysis and interpretation of patient-reported outcomes*. 2nd ed. Chichester: John Wiley.

Fayers, P. and D. Machin (2016). *Quality of Life: The assessment, analysis and reporting of patient-reported outcomes*. Third Edition ed. Chichester, West Sussex: John Wiley and Sons.

Felce, D. and J. Perry (1995). Quality of life: its definition and measurement. *Res Dev Disabil* 16(1) 51-74.

Ferrans, C. E. (1990). Quality of life: conceptual issues. *Semin Oncol Nurs* 6(4) 248-254.

Field, A. (2009). *Discovering statistics using SPSS*. 3rd ed ed. Los Angeles, California: Sage.

Finch, A. P., J. E. Brazier and C. Mukuria (2018). What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *Eur J Health Econ* 19(4) 557-570.

Finch, A. P., et al. (2017). An Exploratory Study on Using Principal-Component Analysis and Confirmatory Factor Analysis to Identify Bolt-On Dimensions: The EQ-5D Case Study. *Value Health* 20(10) 1362-1375.

Fleishman, J. A. and W. F. Lawrence (2003). Demographic variation in SF-12 scores: true differences or differential item functioning? *Medical Care* 41(7 Suppl) III75-III86.

Flynn, T. N. C., Phil Coast, Joanna Peters, Tim J. (2011). Assessing quality of life among British older people using the ICEPOP CAPability (ICECAP-O) measure. *Applied Health Economics And Health Policy* 9(5) 317-329.

Fried, L. P., et al. (2001). Frailty in older adults: evidence for a phenotype. *J Gerontol A Biol Sci Med Sci* 56(3) M146-156.

Fries, J. F., B. Bruce and D. Cella (2005). The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clin Exp Rheumatol* 23(5 Suppl 39) S53-57.

Gage, H., et al. (2014). Specialist rehabilitation for people with parkinson's disease in the community: a randomised controlled trial. *Health Services and Delivery Research* 2.51.

Gale, C., C. Cooper and A. Sayer (2015). Prevalence of frailty and disability: findings from the English Longitudinal Study of Ageing. *Age Ageing* 44(1) 162-165.

Goldberg, D. and P. Williams (1988). *A users guide to the general health questionnaire*. Windsor: NFER-Nelson.

Goldsmith, S. B. (1972). The status of health status indicators. *Health Serv Rep* 87(3) 212-220.

Goodwin, N., et al. (2014). Providing integrated care for older people with complex needs: Lessons from seven international case studies. *The Kings Fund* 1-28.

Greer, A. (1986). The Measurement of Health in Urban Communities. *Journal of Urban Affairs* 8 9-21.

Grewal, I., et al. (2006). Developing attributes for a generic quality of life measure for older people: preferences or capabilities? *Social Science & Medicine* 62(8) 1891-1901 1811p.

Hackert, M. Q. N., J. V. Exel and W. B. F. Brouwer (2017). Valid Outcome Measures in Care for Older People: Comparing the ASCOT and the ICECAP-O. *Value Health* 20(7) 936-944.

Hambleton, R. and R. Jones (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice* 38-47.

Hammar, T., P. Rissanen and M. L. Perälä (2009). The cost-effectiveness of integrated home care and discharge practice for home care patients. *Health Policy* 92(1) 10-20 11p.

Harris, R., et al. (2005). The effectiveness, acceptability and costs of a hospital-at-home service compared with acute hospital care: a randomized controlled trial. *J Health Serv Res Policy* 10(3) 158-166.

Hays, R. and B. Reeve (2010). Measurement and modeling of health-related quality of life. In J. Killewo, H. Heggenhougen and S. Quah eds. *Epidemiology and demography in public health*. San Diego, Academic Press. 195-205.

Hays, R., M. Staquet and P. Fayers (1998). *Quality of life assessment in clinical trials : methods and practice*. Oxford: Oxford University Press.

Hays, R. D., L. S. Morales and S. P. Reise (2000). Item response theory and health outcomes measurement in the 21st century. *Med Care* 38(9 Suppl) II28-42.

Haywood, K. L., et al. (2017). Patient-reported outcome measures in older people with hip fracture: a systematic review of quality and acceptability. *Quality Of Life Research*:

An International Journal Of Quality Of Life Aspects Of Treatment, Care And Rehabilitation 26(4) 799-812.

Haywood, K. L., S. M. Collin and E. Crawley (2014). Assessing severity of illness and outcomes of treatment in children with Chronic Fatigue Syndrome/Myalgic Encephalomyelitis (CFS/ME): a systematic review of patient-reported outcome measures (PROMs). *Child Care Health Dev* 40(6) 806-824.

Haywood, K. L., A. M. Garratt and R. Fitzpatrick (2005). Quality of life in older people: a structured review of generic self-assessed health instruments. *Qual Life Res* 14(7) 1651-1668.

Health and Social Care Information Centre (2014). Health Survey for England 2014: Volume 1 Health, social care and lifestyles. Health and Social Care Information Centre.

Health and social care information centre (2014). Personal social services: expenditure and unit costs England 2013/14.

Health and Social Care Information Centre (2015). Personal Social Services Adult Social Care Survey, England 2014-15.

Health and Social Care Information Centre and Department of Health (2014). Health Survey for England.

Healthwatch Wakefield (2016). Interim report on care home resident interviews: Phase one implementation - care homes not in Vanguard.

Henderson, C., et al. (2013). Cost effectiveness of telehealth for patients with long term conditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. *BMJ (Clinical Research Ed.)* 346 f1035-f1035.

Herdman, M., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 20(10) 1727-1736.

Hicks, S., L. Tinkler and P. Allin (2013). Measuring subjective well-being and its potential role in policy: Perspectives for the UK Office for National Statistics. *Social Indic Res*.

Hofman, C. S., et al. (2017). Examining the construct and known-group validity of a composite endpoint for The Older Persons and Informal Caregivers Survey Minimum Data Set (TOPICS-MDS); A large-scale data sharing initiative. *Plos One* 12(3) e0173081-e0173081.

Holland, R., et al. (2004). Assessing quality of life in the elderly: a direct comparison of the EQ-5D and AQoL. *Health Econ* 13(8) 793-805.

Hospital Episode Statistics Analysis and Health and Social Care Information Centre (2015). Hospital Episode Statistics - Admitted Patient Care, England - 2014-15.

Hulme, C., et al. (2004). Using the EQ-5D to assess health-related quality of life in older people. *Age Ageing* 33(5) 504-507.

Hultberg, E. L., K. Lönnroth and P. Allebeck (2005). Interdisciplinary collaboration between primary care, social insurance and social services in the rehabilitation of people with musculoskeletal disorder: effects on self-rated health and physical performance. *J Interprof Care* 19(2) 115-124.

Hultberg, E. L., K. Lönnroth and P. Allebeck (2007). Effects of a co-financed interdisciplinary collaboration model in primary health care on service utilisation among patients with musculoskeletal disorders. *Work* 28(3) 239-247.

Humphries, R. (2015). Integrated health and social care in England – Progress and prospects. *Health Policy* 119(7) 856-859.

Hunt, S. M., et al. (1980). A quantitative approach to perceived health status: a validation study. *J Epidemiol Community Health* 34(4) 281-286.

Hunt, S. M., et al. (1981). The Nottingham Health Profile: subjective health status and medical consultations. *Soc Sci Med A* 15(3 Pt 1) 221-229.

Jakobsson, U. (2007). Using the 12-item Short Form health survey (SF-12) to measure quality of life among older people. *Aging Clin Exp Res* 19(6) 457-464.

Jakobsson, U., et al. (2012). Construct validity of the SF-12 in three different samples. *Journal Of Evaluation In Clinical Practice* 18(3) 560-566.

Janssen, B. and A. Szende (2014). Population Norms for the EQ-5D. In A. Szende, B. Janssen and J. Cabases eds. *Self-Reported Population Health: An International Perspective based on EQ-5D*. Dordrecht, Springer.

Janssen, M. F., et al. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res* 22(7) 1717-1727.

Jones, K., et al. (2013). Personalization in the health care system: do personal health budgets have an impact on outcomes and cost? *Journal Of Health Services Research & Policy* 18(2 Suppl) 59-67.

Kaambwa, B., et al. (2015). An empirical comparison of the OPQoL-Brief, EQ-5D-3 L and ASCOT in a community dwelling population of older people. *Health Qual Life Outcomes* 13 164.

Kahneman, D., P. Wakker and R. Sarin (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics* 112(2) 375-406.

Kaiser, H. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*(20) 141-151.

Kammann, R. and R. Flett (1983). A scale to measure current level of general happiness. *Aust Psychol* 35 259-265.

Karimi, M. and J. Brazier (2016). Health, Health-Related Quality of Life, and Quality of Life: What is the Difference? *Pharmacoeconomics* 34(7) 645-649.

Keetharuth, A. D., et al. (2018). Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. *Br J Psychiatry* 212(1) 42-49.

Knafl, K., et al. (2007). The analysis and interpretation of cognitive interviews for instrument development. *Res Nurs Health* 30(2) 224-234.

Knott, R. J., et al. (2017). Differential item functioning in quality of life measurement: An analysis using anchoring vignettes. *Soc Sci Med* 190 247-255.

Kodner, D. and C. Spreeuwenberg (2002). Integrated care : meaning, logic, applications, and implications – a discussion paper. *International Journal Of Integrated Care* 2(November) 1-6.

Kunz, S. (2010). Psychometric properties of the EQ-5D in a study of people with mild to moderate dementia. *Qual Life Res* 19(3) 425-434.

Kuyken, W. and T. Group (1995). The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization. *Social Science and Medicine* 41 1403-1409.

Larson, J. (1991). *The Measurement of Health: Concepts and Indicators*. Westport, CT: Greenwood.

Larson, J. S. (1999). The conceptualization of health. *Med Care Res Rev* 56(2) 123-136.

Liang, Y. and W. Wu (2014). Exploratory analysis of health-related quality of life among the empty-nest elderly in rural China: an empirical study in three economically developed cities in eastern China. *Health Qual Life Outcomes* 12 59.

Lix, L. M., et al. (2016). Differential Item Functioning in the SF-36 Physical Functioning and Mental Health Sub-Scales: A Population-Based Investigation in the Canadian Multicentre Osteoporosis Study. *PLoS One* 11(3) e0151519.

Lumley, J., et al. (2006). PRISM (Program of Resources, Information and Support for Mothers): a community-randomised trial to reduce depression and improve women's physical health six months after birth [ISRCTN03464021]. *BMC Public Health* 6 37.

Lung, T., et al. (2017) Comparison of the HUI3 and the EQ-5D-3L in a nursing home setting. *Plos one* [online]. Available at: <http://onlinelibrary.wiley.com/o/cochrane/clcentral/articles/272/CN-01341272/frame.html>.

Lutomski, J. E., et al. (2017). Measurement properties of the EQ-5D across four major geriatric conditions: Findings from TOPICS-MDS. *Health And Quality Of Life Outcomes* 15(1) 45-45.

Makai, P., et al. (2014). Quality of life instruments for economic evaluations in health and social care for older people: A systematic review. *Social Science & Medicine* 102 83-93 11p.

Malley, J. N., et al. (2012). An assessment of the construct validity of the ASCOT measure of social care-related quality of life with older people. *Health And Quality Of Life Outcomes* 10 21-21.

Mallinson, S. (2002). Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. *Soc Sci Med* 54(1) 11-21.

Maruish, M. E. E. (2012). *User's manual for the SF-12v2 Health Survey*. 3rd Edition ed. Lincoln, RI: QualityMetric Incorporated.

McDonald, R. (1999). *Test Theory: A Unified Treatment*. 1st Edition ed: Psychology Press.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *J Appl Psychol* 95(4) 728-743.

Milfont, T. and R. Fischer (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research* 3 111-121.

Milte, C. M., et al. (2014). How important is health status in defining quality of life for older people? An exploratory study of the views of older South Australians. *Applied Health Economics And Health Policy* 12(1) 73-84.

Mokkink, L. B., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 63(7) 737-745.

Moser, D. K., et al. (2013). 'It could be worse ... lot's worse!' Why health-related quality of life is better in older compared with younger individuals with heart failure. *Age Ageing* 42(5) 626-632.

Mukuria, C., et al. (2016). An empirical comparison of well-being measures used in UK. Policy Research Unit in Economic Evaluation of Health and Social Care Interventions.

Muthén, L. K. and B. O. Muthén (1998-2015). *Mplus User's Guide*. Seventh Edition ed. Los Angeles, CA: Muthén & Muthén.

Nasreddine, Z. S., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 53(4) 695-699.

National Institute for Health and Care Excellence (2013). Guide to the methods of technology appraisal 2013. 1-93.

National Institute for Health and Care Excellence (2016). The Social Care Guidance Manual.

National Institute for Health and Clinical Excellence (2013). The Social Care Guidance Manual.

National Institute for Health Research. (2014). *The Community Ageing Research 75+ (CARE 75+) cohort study [online]*. Available at: <http://clahrc-yh.nihr.ac.uk/our-themes/primary-care-based-management-of-frailty-in-older-people/projects/the-yorkshire-humber-community-ageing-research-care-study> [Accessed 7th July 2018].

Navarro, V. (1977). *Health and Medical Care in the U.S.: A Critical Analysis*. Farmingdale, NY: Baywood.

Neilson, E. (1988). Health Values: Achieving High Level Wellness-Origins, Philosophy, Purpose. *Health Values* 12 3-5.

Netten, A. (2011). Overview of outcome measurement for adults using social care services and support.

Netten, A., et al. (2012). Outcomes of social care for adults: developing a preference-weighted measure. *Health Technology Assessment* 16(50) 1-166 166p.

Nguyen, T. H., et al. (2014). An introduction to item response theory for patient-reported outcome measurement. *Patient* 7(1) 23-35.

NHS Benchmarking Network (2016). Older People's Care in Acute Settings: Benchmarking Report.

NHS Digital. (2015). *Personal Social Services Adult Social Care Survey, England 2014-15 [online]*. Available at: <https://digital.nhs.uk/data-and-information/publications/statistical/personal-social-services-adult-social-care-survey/personal-social-services-adult-social-care-survey-england-2014-15> [Accessed 19th November 2018].

Nicholson, C., et al. (2017). Supportive care for older people with frailty in hospital: An integrative review. *Int J Nurs Stud* 66 60-71.

Noack, H. (1987). Concepts of Health and Health Promotion. In T. Abelin, Z. Brzezinski and V. Carstairs eds. *Measurement in Health Promotion and Protection*. Copenhagen, Denmark, World Health Organisation. 5-28.

Office for National Statistics (2012). Summary of results from testing of experimental subjective well-being questions. Office for National Statistics.

Office for National Statistics (2013). General Lifestyle Survey: 2011.

Office for National Statistics (2016a). Disability-Free Life Expectancy (DFLE) and Life Expectancy (LE): at age 65 by region, England.

Office for National Statistics (2016b). National life tables, UK: 2013 to 2015.

Office for National Statistics (2016c). Quality and Methodology Information - Personal Well-being in the UK (Annual and Three-year estimates). ONS.

Office for National Statistics (2017a). Families and Households: 2017.

Office for National Statistics (2017b). National Population Projections: 2016-based

Office for National Statistics (2018). Population pyramids, 1966, 2016 and 2066 (principal projection), UK (Source: Population estimates, Principal population projections, 2016-based, Office for National Statistics). Living longer - how our population is changing and why it matters.

Office of National Statistics (2014). Life Expectancy at Birth and at Age 65 by Local Areas in the United Kingdom: 2006-08 to 2010-12.

Oguz, S., et al. (2013). Measuring National Well-being - What matters most for personal well-being?, Office for National Statistics.

Pannenborg, C. (1979). *A New International Health Order: An Inquiry into the International Relations of World Health and Medical Care*. The Netherlands: Sijthoff and Noordhoff.

Parsons, N., et al. (2014). Outcome assessment after hip fracture: is EQ-5D the answer? *Bone Joint Res* 3(3) 69-75.

Parsons, T. (1972). In E. Jaco ed. *Patients, Physicians and Illness*. New York, Free Press.

Patrick, D. L., et al. (2011a). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product

evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. *Value Health* 14(8) 967-977.

Patrick, D. L., et al. (2011b). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. *Value Health* 14(8) 978-988.

Peasgood, T., et al. (2014). A conceptual comparison of well-being measures used in the UK. Policy Research Unit in Economic Evaluation of Health and Care Interventions.

Petrillo, J., et al. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health* 18(1) 25-34.

Pettit, T., et al. (2001). Validation and normative data of health status measures in older people: the Islington study. *International Journal Of Geriatric Psychiatry* 16(11) 1061-1070.

PricewaterhouseCoopers and Australian Government Department of Health and Ageing (2007). National Evaluation of the Second Round of Coordinated Care Trials.

Priede, C. and S. Farrall (2011). Comparing results from different styles of cognitive interviewing: 'verbal probing' vs. 'thinking aloud'. *International Journal of Social Research Methodology* 14(4) 271-287.

Putnick, D. L. and M. H. Bornstein (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Dev Rev* 41 71-90.

QSR International Pty Ltd (2013). NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

Ralph, K., K. Palmer and J. Olney (2011). Subjective Well-being: A qualitative investigation of subjective well-being questions. Office of National Statistics: A working paper for the Technical Advisory Group.

RAND. (2015a). *12-Item Short Form Survey (SF-12)* [online]. Available at: http://www.rand.org/health/surveys_tools/mos/12-item-short-form.html [Accessed 5th September 2016].

RAND. (2015b). *20-Item Short Form Survey (SF-20)* [online]. Available at: http://www.rand.org/health/surveys_tools/mos/20-item-short-form.html.

Ratcliffe, J., et al. (2017). An Empirical Comparison of the EQ-5D-5L, DEMQOL-U and DEMQOL-Proxy-U in a Post-Hospitalisation Population of Frail Older People Living in Residential Aged Care. *Applied Health Economics And Health Policy* 15(3) 399-412.

Ratcliffe, J., et al. (2017). Does one size fit all? Assessing the preferences of older and younger people for attributes of quality of life. *Qual Life Res* 26(2) 299-309.

Reeve, B. B., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 45(5 Suppl 1) S22-31.

Reid, G. e., al (2007). Change and transformation: the impact of an action-research evaluation on the development of a new service. *Learning in Health and Social Care* 6(2) 61-71.

Reise, S. P., W. E. Bonifay and M. G. Haviland (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *J Pers Assess* 95(2) 129-140.

Rejeski, W. J. and S. L. Mihalko (2001). Physical activity and quality of life in older adults. *J Gerontol A Biol Sci Med Sci* 56 Spec No 2 23-35.

Resnick, B. and E. S. Nahm (2001). Reliability and validity testing of the revised 12-item Short-Form Health Survey in older adults. *Journal Of Nursing Measurement* 9(2) 151-161.

Resnick, B. and R. Parker (2001). Simplified scoring and psychometrics of the revised 12-item Short-Form Health Survey. *Outcomes Management For Nursing Practice* 5(4) 161-166.

Robertson, C., et al. (2009). Meaning behind measurement: self-comparisons affect responses to health-related quality of life questionnaires. *Qual Life Res* 18(2) 221-230.

Robertson, H. (2011). Integration of health and social care: A review of literature and models. Implications for Scotland.

Robeyns, I. (2003). Sen's capability approach and gender inequality: Selecting relevant capabilities. *Feminist Economics* 9((2-3)) 61-92.

Rockwood, K., et al. (2006). Long-term risks of death and institutionalization of elderly people in relation to deficit accumulation at age 70. *J Am Geriatr Soc* 54(6) 975-979.

Rothman, M., et al. (2009). Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. *Value Health* 12(8) 1075-1083.

Rothrock, N. E., K. A. Kaiser and D. Cella (2011). Developing a valid patient-reported outcome measure. *Clin Pharmacol Ther* 90(5) 737-742.

Sahota , O., et al. (2016). Comparing the cost-effectiveness and clinical effectiveness of a new community in-reach rehabilitation service with the cost-effectiveness and clinical effectiveness of an established hospital-based rehabilitation service for older people. *Health Services and Delivery Research* 4(7).

Samejima, F. (1996). The graded response model. In W. van der Linden and R. Hambleton eds. *Handbook of Modern Item Response Theory*. New York, NY, Springer.

Sanchez-Arenas, R., et al. (2014). Value of EQ-5D in Mexican city older population with and without dementia (SADEM study). *Int J Geriatr Psychiatry* 29(5) 478-488.

School of Health and Related Research. (2018). *Extending the QALY* [online]. Available at: <https://scharr.dept.shef.ac.uk/e-qaly/about-the-project/> [Accessed 26th November 2018].

Schroeder, E. (1983). Concepts of Health and Illness. In A. Culyer ed. *Health Indicators: An International Study for the European Science Foundation*. New York, St Martins. 23-33.

Schwarz, N. and G. Clore (2004). Mood as information: 20 years later. *Psychological Inquiry* 14 296-303.

Sen, A. (1982). *Choice, welfare and measurement*. Cambridge: MA: Harvard University Press.

Sen, A. (1983). Capability and well-being. In M. Nussbaum ed. *The quality of life*. Oxford, Clarendon Press.

Sheffield Clinical Commissioning Group (2015). People Keeping Well in their Community Briefings July 2015.

Sheffield Teaching Hospitals NHS Foundation Trust. (2019). *The Online Public Advisory Panel* [online]. Available at: <https://www.sheffieldclinicalresearch.org/for-patients-public/how-to-get-involved/online-public-advisory-panel/> [Accessed 12 January 2019].

Shou, J., et al. (2016). Reliability and validity of 12-item Short-Form health survey (SF-12) for the health status of Chinese community elderly population in Xujiahui district of Shanghai. *Aging Clinical & Experimental Research* 28(2) 339-346.

Smith, A. B., et al. (2016). A Differential Item Functioning Analysis of the EQ-5D in Cancer. *Value Health* 19(8) 1063-1067.

Sommers, L. S., et al. (2000). Physician, nurse, and social worker collaboration in primary care for chronically ill seniors. *Arch Intern Med* 160(12) 1825-1833.

Speller, V., A. Learmonth and D. Harrison (1997). The search for evidence of effective health promotion. *BMJ* 315(7104) 361-363.

Sprangers, M. A. and C. E. Schwartz (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med* 48(11) 1507-1515.

Spuling, S., J. Wolff and S. Wurm (2017). Response shift in self-rated health after serious health events in old age. *Social science and medicine* 192 85-93.

StataCorp. (2015). Stata Statistical Software: Release 14. College Station, TX, StataCorp LP.

Statistical Consulting Group, Institute for Digital Research and Education and UCLA Stata2Mplus.

Steinberg, L. and Thissen, D. (1996) Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, 1(1), 81-97.

Stewart-Brown, S., et al. (2009). Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. *Health Qual Life Outcomes* 7 15.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess* 80(1) 99-103.

Sulch, D., et al. (2002). Integrated care pathways and quality of life on a stroke rehabilitation unit. *Stroke; A Journal Of Cerebral Circulation* 33(6) 1600-1604.

Sulch, D., et al. (2000). Randomized controlled trial of integrated (managed) care pathway for stroke rehabilitation. *Stroke* 31(8) 1929-1934.

Taggart, F., et al. (2013). Cross cultural evaluation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) --a mixed methods study. *Health Qual Life Outcomes* 11 27.

Tay, L., A. Meade and M. Cao (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods* 18(1) 3-46.

Tennant, R., et al. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health Qual Life Outcomes* 5 63.

Teresi, J. A., et al. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Qual Life Res* 16 Suppl 1 43-68.

Teresi, J. A., et al. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychol Sci Q* 51(2) 148-180.

Terwee, C. B., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 60(1) 34-42.

Terwee, C. B., et al. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 21(4) 651-657.

The Better Care Fund (2014). Better Care Fund Policy Framework.

Tidermark, J. and G. Bergström (2007). Responsiveness of the EuroQol (EQ-5D) and the Nottingham Health Profile (NHP) in elderly patients with femoral neck fractures. *Qual Life Res* 16(2) 321-330.

Tidermark, J., et al. (2003). Responsiveness of the EuroQol (EQ 5-D) and the SF-36 in elderly patients with displaced femoral neck fractures. *Qual Life Res* 12(8) 1069-1079.

Tidermark, J., et al. (2002). Femoral neck fractures in the elderly: functional outcome and quality of life according to EuroQol. *Qual Life Res* 11(5) 473-481.

Tinkler, L. and S. Hicks (2011). Supplementary Paper: Measuring Subjective Well-being. Office for National Statistics.

Torrance, G. W. (1987). Utility approach to measuring health-related quality of life. *J Chronic Dis* 40(6) 593-603.

Torrance, G. W., W. H. Thomas and D. L. Sackett (1972). A utility maximization model for evaluation of health care programs. *Health Serv Res* 7(2) 118-133.

Toseland, R. W., et al. (1996). Outpatient geriatric evaluation and management. Results of a randomized trial. *Medical Care* 34(6) 624-640.

Tourangeau, R. (1984). *Cognitive science and survey methods*. Washington, DC: National Academy Press.

Tourangeau, R., L. Rips and K. Rasinski (2000). *The pyschology of survey response*. Cambridge: The Cambridge University Press.

Turner, G., et al. (2014). Best practice guidelines for the management of frailty: a British Geriatrics Society, Age UK and Royal College of General Practitioners report. *Age Ageing* 43(6) 744-747.

U.S National Library of Medicine (1993-2018). ClinicalTrials.gov.

Ubel, P., et al. (2005). What is perfect health to an 85-year-old?: evidence for scale recalibration in subjective health ratings. *Medical Care* 43(10) 1054-1057.

van Hout, B., et al. (2012). Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health* 15(5) 708-715.

van Leeuwen, K. M., et al. (2015a). Dutch translation and cross-cultural validation of the Adult Social Care Outcomes Toolkit (ASCOT). *Health Qual Life Outcomes* 13 56.

van Leeuwen, K. M., et al. (2015b). Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. *Value In Health: The Journal Of The International Society For Pharmacoeconomics And Outcomes Research* 18(1) 35-43.

van Leeuwen, K. M., et al. (2015). Exploration of the content validity and feasibility of the EQ-5D-3L, ICECAP-O and ASCOT in older adults. *BMC Health Serv Res* 15 201.

van Leeuwen, K. M., et al. (2014). What can local authorities do to improve the social care-related quality of life of older adults living at home? Evidence from the Adult Social Care Survey. *Health & Place* 29 104-113.

Vandenberg, R. and C. Lance (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 2 4-69

Verbrugge, L. M. and A. M. Jette (1994). The disablement process. *Soc Sci Med* 38(1) 1-14.

Ware, J. E., J. B. Bjorner and M. Kosinski (2000). Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Med Care* 38(9 Suppl) II73-82.

Ware, J. E., et al. (1981). Choosing measures of health status for individuals in general populations. *Am J Public Health* 71(6) 620-625.

Ware, J. E., et al. (1993). *SF-36 health survey: manual and interpretation guide*: The Health Institute, New England Medical Center.

WEMWBS Research Team. *Frequent Issues in Translation [online]*. Available at: https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs/researchers/languages/frequent_issues_in_translation.pdf [Accessed 8th October 2018].

What Works for Wellbeing. (2018). *What Works for Wellbeing [online]*. Available at: <https://whatworkswellbeing.org/> [Accessed 18th December 2018].

Wiggins, R., et al. (2008). The evaluation of a self-enumerated scale of quality of life (CASP-19) in the context of research on ageing: A combination of exploratory and confirmatory approaches. *Social Indicators Research* 89(1) 61-77.

Williams, H. A. (1993). A comparison of social support and social networks of black parents and white parents with chronically ill children. *Soc Sci Med* 37(12) 1509-1520.

Willis, G. (2005). *Cognitive Interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: SAGE Publications.

Windle, K. e., al (2009). *National evaluation of Partnerships for older people projects: final report*. Canterbury: Personal Social Services Research Unit.

Wood, P. (1986). Health and Disease and Its Importance for Models Relevant to Health Research. In B. Nizetic, H. Pauli and P. Svensson eds. *Scientific Approaches to Health and Health Care*. Copenhagen, Denmark, World Health Organisation.

World Health Organization (2002). Proposed working definition of an older person in Africa for the MDS Project.

World Health Organization (2015). *First WHO ministerial conference on global action against dementia: 16-17 March 2015, Geneva, Switzerland: meeting report*. Geneva: World Health Organization.

Yu, C. (2002). Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes. Los Angeles, University of California.

Yu, Y. F., A. P. Yu and J. Ahn (2007). Investigating differential item functioning by chronic diseases in the SF-36 health survey: a latent trait analysis using MIMIC models. *Med Care* 45(9) 851-859.

Appendices

Chapter 3 appendices

Appendix 1 – Search strategy for rapid review of PROMs in integrated health and social care evaluations in MEDLINE

Search	
#1	“Health related quality of life” OR “HRQoL” OR “Social care related quality of life” OR “SCRQoL” OR “QoL” OR “quality of life” OR “Wellbeing” OR “Well-being” OR “Preference based” OR “Social value” OR “Social impact” OR “Social capital”
#2	“Integration” OR “Integrated care” OR “Pooled budgets” OR “Integrated working” OR “Joint working” OR “Integrated budgets” OR “Inter-disciplinary” OR “Interdisciplinary” OR “Multi-disciplinary” OR “multidisciplinary” OR “Cross-sector” OR “Cross-sectoral” OR “Evaluation” OR “Economic evaluation” OR “Cost-effectiveness” OR “Cost-utility analysis” OR “Social return on investment” OR “SROI”
#3	“Health care” OR “Healthcare” OR “Health”
#4	“social care” OR “long term care”
#5	#3 and #4
#6	#1 and #2 and #5

Appendix 2 – Included integration schemes and generic PROMs used

Study	Scheme	EQ-5D	SF-36	SF-20	SF-12	ASCOT	NHP	ICECAP-O
Anderson et al. 2000	Hospital at home stroke		1				1	
Ariss et al. 2015	Community rehab and intermediate care	1						
Cartwright et al. 2013/ Henderson et al. 2013	Telehealth WSD trial	1			1			1
Gage et al. 2014	Community rehab Parkinson's	1	1					
Hammar et al. 2009	Integrated home care and discharge	1					1	
Harris et al. 2005	Hospital at home		1					
Hultberg et al. 2005/ Hultberg et al. 2007	DELTA MDTs Sweden	1						
Jones et al. 2013	Personal health budgets	1				1		
Lumley et al. 2006	PRISM		1					
PWC 2007	CCT1			1				
PWC 2007	CCT2	1						
Reid et al. 2007	Care management rehab link teams	1						
Sahota et al. 2016	Community rehab	1						
Sommers et al. 2000	Physician nurse social worker collaboration		1					
Sulch et al. 2002/ Sulch et al. 2000	ICP stroke	1						
Toseland et al. 1996	Outpatient geriatric evaluation and management			1				
Windle et al. 2009	POPP	1						
Total		11	6	1	1	1	2	1

PWC= PricewaterhouseCoopers H+SC= Health and social care POPP= partnership for older people projects ED= emergency department MDTs= multidisciplinary teams MHT= mental health team MH= mental health ICP= integrated care pathway COPD = chronic obstructive pulmonary disease PACE= program of all-inclusive care for the elderly PRISMA= program of resources to integrate services for the maintenance of autonomy SIPA= system of integrated care for older persons CCT= coordinated care trial S/HMO= social health maintenance organisation PRISM= program of resources, information and support for mothers WSD= Whole systems demonstrator

Appendix 3 – Search strategy systematic review of psychometric evidence of included PROMs in MEDLINE

Search	Search Terms
#1	"Valid*" OR "Accept*" OR "Feas*" OR "Develop*" OR "Reliab*" OR "Measure* properties" OR "measure* performance" OR "Psychometric*" OR "Item response theory" OR "Rasch" OR "IRT" OR "Differential item functioning" OR "DIF" OR "Measurement invariance"
#2	"Elder*" OR "Old*" OR "Frail*"
#3	"EQ-5D" OR "EQ-5D-3L" OR "EQ-5D-5L" OR "Euroqol" OR "WEMWBS" OR "SWEMWBS" OR "Warwick Edinburgh mental well* scale" OR "Short Warwick Edinburgh mental well* scale" OR "ONS4" OR "ONS-4" OR "ONS subjective well*" OR "ONS personal well*" OR "SF-12" OR "ASCOT" OR "Adult social care outcomes toolkit"
#4	#1 AND #2 AND #3

Appendix 4 – COSMIN checklist

COSMIN checklist with 4-point scale

Contact
CB Terwee, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl, www.emgo.nl
E-mail: cb.terwee@vumc.nl



Instructions

This version of the COSMIN checklist is recommended for use in systematic reviews of measurement properties. With this version it is possible to calculate overall methodological quality scores per study on a measurement property. A methodological quality score per box is obtained by taking the lowest rating of any item in a box ('worse score counts'). For example, if for a reliability study one item in the box 'Reliability' is scored poor, the methodological quality of that reliability study is rated as poor. The Interpretability box and the Generalizability box are mainly used as data extraction forms. We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box (e.g. norm scores, floor-ceiling effects, minimal important change) of the instruments under study from the included articles. Similar, we recommend to use the Generalizability box to extract data on the characteristics of the study population and sampling procedure. Therefore no scoring system was developed for these boxes.

This scoring system is described in this paper:

Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research 2011, July 6 [epub ahead of print].

Step 1. Evaluated measurement properties in the article

	Internal consistency	Box A
	Reliability	Box B
	Measurement error	Box C
	Content validity	Box D
	Structural validity	Box E
	Hypotheses testing	Box F
	Cross-cultural validity	Box G
	Criterion validity	Box H
	Responsiveness	Box I

Step 2. Determining if the statistical method used in the article are based on CTT or IRT

Box General requirements for studies that applied Item Response Theory (IRT) models		excellent	good	fair	poor
1 Was the IRT model used adequately described? e.g. One Parameter Logistic Model (OPLM), Partial Credit Model (PCM), Graded Response Model (GRM)	IRT model adequately described	IRT model not adequately described			
2 Was the computer software package used adequately described? e.g. RUMM2020, WINSTEPS, OPLM, MULTILOG, PARSCALE, BILOG, NLMIXED	Software package adequately described	Software package not adequately described			
3 Was the method of estimation used adequately described? e.g. conditional maximum likelihood (CML), marginal maximum likelihood (MML)	Method of estimation adequately described	Method of estimation not adequately described			
4 Were the assumptions for estimating parameters of the IRT model checked? e.g. unidimensionality, local independence, and item fit (e.g. differential item functioning (DIF))	assumptions of the IRT model checked	assumptions of the IRT model partly checked	assumptions of the IRT model not checked or unknown		

To obtain a total score for the methodological quality of studies that use IRT methods, the 'worse score counts' algorithm should be applied to the IRT box in combination with the box of the measurement property that was evaluated in the IRT study. For example, if IRT methods are used to study internal consistency and item 4 in the IRT box is scored fair, while the items in the internal consistency box (box A) are all scored as good or excellent, the methodological quality score for internal consistency will be fair. However, if any of the items in box A is scored poor, the methodological quality score for internal consistency will be poor.

Step 3. Determining if a study meets the standards for good methodological quality

Box A. Internal consistency		excellent	good	fair	poor
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model? <i>Design requirements</i>					
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described			
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled		
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)	
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population		Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥ 100	5* #items and ≥ 100 OR 6-7* #items but <100	5* #items but <100		<5* #items

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	good	fair	poor
<i>Design requirements</i>				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
<i>Statistical methods</i>				
9 for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
9 for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
11 for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2 , reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

		Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions				
Statistical methods					
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box C. Measurement error: absolute measures

		excellent	good	fair	poor
<i>Design requirements</i>					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Were at least two measurements available?	At least two measurements			Only one measurement
5	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6	Was the time interval stated?	Time interval stated		Time interval NOT stated	
7	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
<i>Statistical methods</i>				
11 for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population

Box D. Content validity (including face validity)

	excellent	good	fair	poor
<i>General requirements</i>				
1 Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured	Aspects of the construct to be measured poorly described AND this was not taken into consideration		NOT assessed if all items refer to relevant aspects of the construct to be measured

2 Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥ 10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3 Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4 Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehensively reflect the construct to be measured
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box E. Structural validity

	excellent	good	fair	poor
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model? <i>Design requirements</i>				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the analysis adequate?	7* #items and ≥ 100	5* #items and ≥ 100 OR 5-7* #items but <100	5* #items but <100	$<5*$ #items
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. rotation method not described)	Other important methodological flaws in the design or execution of the study (e.g. inappropriate rotation method)

Statistical methods

6 for CTT: Was exploratory or confirmatory factor analysis performed?	Exploratory or confirmatory factor analysis performed and type of factor analysis appropriate in view of existing information	Exploratory factor analysis performed while confirmatory would have been more appropriate	No exploratory or confirmatory factor analysis performed
7 for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	IRT test for determining (uni)dimensionality performed		IRT test for determining (uni)dimensionality NOT performed

Box F. Hypotheses testing

	excellent	good	fair	Poor
<i>Design requirements</i>				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)

4 Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5 Was the expected <i>direction</i> of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6 Was the expected absolute or relative <i>magnitude</i> of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7 for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8 for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
9 Were there any important flaws in the design or methods of the study? <i>Statistical methods</i>	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study	
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box G. Cross-cultural validity				
	excellent	good	fair	poor
<i>Design requirements</i>				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

3 Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4 Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5 Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6 Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7 Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8 Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		
9 Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10 Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11 Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12 Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

<i>Statistical methods</i>			
14 for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed		Multiple-group confirmatory factor analysis NOT performed
15 for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed		DIF between language groups NOT assessed

Box H. Criterion validity		excellent	good	fair	poor
<i>Design requirements</i>					
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described			
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled		
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)	
4 Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'	

5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
<i>Statistical methods</i>			
6 for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated		Correlations or AUC NOT calculated
7 for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated		Sensitivity and specificity NOT calculated

Box I. Responsiveness		excellent	good	fair	poor
<i>Design requirements</i>					
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described			
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled		
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)	
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used	
5 Was the time interval stated?	Time interval adequately described			Time interval NOT described	

6 If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7 Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
<i>Design requirements for hypotheses testing</i>				
For constructs for which a gold standard was not available:				
8 Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected	
9 Was the expected <i>direction</i> of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10 Were the expected absolute or relative <i>magnitude</i> of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

11 Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)	
12 Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	
13 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study	
<i>Statistical methods</i>				
14 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate	

<i>Design requirement for comparison to a gold standard</i>					
For constructs for which a gold standard was available:					
15	Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	
<i>Statistical methods</i>					
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Interpretability

We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box of the instruments under study from the included articles.

Box Interpretability	
Percentage of missing items	
Description of how missing items were handled	
Distribution of the (total) scores	
Percentage of the respondents who had the lowest possible (total) score	
Percentage of the respondents who had the highest possible (total) score	
Scores and change scores (i.e. means and SD) for relevant (sub) groups, e.g. for normative groups, subgroups of patients, or the general population	
Minimal Important Change (MIC) or Minimal Important Difference (MID)	

Generalizability

We recommend to use the Generalizability box to extract data on the characteristics of the study populations and sampling procedures of the included studies.

Box Generalisability	
Median or mean age (with standard deviation or range)	
Distribution of sex	
Important disease characteristics (e.g. severity, status, duration) and description of treatment	
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	
Countries in which the study was conducted	
Language in which the HR-PRO instrument was evaluated	
Method used to select patients (e.g. convenience, consecutive, or random)	
Percentage of missing responses (response rate)	

Appendix 5 – List of included studies

- Bentur, N., King, Y. (2010). The challenge of validating SF-12 for its use with community-dwelling elderly in Israel. *Quality of life research*. 19(1), p91-95.
- Brazier, J. E., Walters, S. J., Nicholl, J. P., Kohler, B. (1996). Using the SF-36 and Euroqol on an elderly population. *Quality of life research*. 5(2), p195-204.
- Cernin, P. A., Cresci, K., Jankowski, T. B., Lichtenberg, P. A. (2010). Reliability and validity testing of the Short-Form Health Survey in a sample of community-dwelling African American older adults. *Journal of Nursing Measurement*. 18(1), p49-59.
- Coast, J., Peters, T. J., Richards, S. H., Gunnell, D. J. (1998). Use of the EuroQoL among elderly acute care patients. *Quality of life research*. 7(1), p1-10.
- Davis, J. C., Bryan, S., McLeod, R., Rogers, J., Khan, K., Liu-Ambrose, T. (2012). Exploration of the association between quality of life, assessed by the EQ-5D and ICECAP-O, and falls risk, cognitive function and daily function, in older adults with mobility impairments. *BMC geriatrics*. 12, p65.
- Diaz-Redondo, A., Rodriguez-Blazquez, C., Ayala, A., Martinez-Martin, P., Forjaz, M. J. (2014). EQ-5D rated by proxy in institutionalized older adults with dementia: psychometric pros and cons. *Geriatrics & Gerontology International*. 14(2), p346-353.
- Fleishman, J., Lawrence, W. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning? *Medical Care*. 41(7 Suppl), pIII75-III86.
- Hackert, M. Q. N., Exel, J. V., Brouwer, W. B. F. (2017). Valid Outcome Measures in Care for Older People: Comparing the ASCOT and the ICECAP-O. *Value in Health*. 20(7), p936-944.
- Holland, R., Smith, R., Harvey, I., Swift, L., Lenaghan, E. (2004). Assessing quality of life in the elderly: a direct comparison of the EQ-5D and AQoL. *Health Economics*. 13(8), p793-805.
- Jakobsson, U. (2007) Using the 12-item short form health survey (SF-12) to measure quality of life among older people. *Aging Clin Exp Res*. 19(6), p457-464.
- Jakobsson, U., Westergren, A., Lindskov, S., Hagell, P. (2012). Construct validity of the SF-12 in three different samples. *Journal of Evaluation in Clinical Practice*. 18(3), p560-566.
- Kaambwa, B., Gill, L., McCaffrey, N., Lancsar, E., Cameron, I. D., et. al. (2015). An empirical comparison of the OPQoL-Brief, EQ-5D-3 L and ASCOT in a community dwelling population of older people. *Health and Quality of Life Outcomes*. 13, p164.

Liang, Y. Wu, W. (2014). Exploratory analysis of health-related quality of life among the empty-nest elderly in rural China: an empirical study in three economically developed cities in eastern China. *Health and Quality Of Life Outcomes*. 12, p59.

Lung, T., Howard, K., Etherton-Bear, C., Sim, M., Lewin, G., Arends, G. (2017). Comparison of the HUI3 and the EQ-5D-3L in a nursing home setting. *Plos One*. 12(2).

Lutomski, J. E., Krabbe, P. F. M., Bleijenberg, N., Blom, J., Kempen, G. I. et. al. (2017). Measurement properties of the EQ-5D across four major geriatric conditions: Findings from TOPICS-MDS. *Health and Quality of Life Outcomes*. 15(1), p45.

Malley, J. N., Towers, A., Netten, A. P., Brazier, J. E., Forder, J. E., Flynn, T. (2012). An assessment of the construct validity of the ASCOT measure of social care-related quality of life with older people. *Health and Quality of Life Outcomes*. 10, p21.

Netten, A., Burge, P., Malley, J., Potoglou, D., Towers, A. M. et. al. (2012). Outcomes of social care for adults: developing a preference-weighted measure. *Health Technology Assessment*. 16, p166.

Parsons, N., Griffin, X. L., Achten, J., Costa, M. L. (2014). Outcome assessment after hip fracture: is EQ-5D the answer? *Bone & Joint Research*. 3(3), p69-75.

Pettit, T., Livingston, G., Manela, M., Kitchen, G., Katona, C., Bowling, A. (2001). Validation and normative data of health status measures in older people: the Islington study. *International Journal of Geriatric Psychiatry*. 16(11), p1061-1070.

Ratcliffe, J., Flint, T., Easton, T., Killington, M., Cameron, I. (2017). An Empirical Comparison of the EQ-5D-5L, DEMQOL-U and DEMQOL-Proxy-U in a Post-Hospitalisation Population of Frail Older People Living in Residential Aged Care. *Applied Health Economics and Health Policy*. 15(3), p399-412.

Resnick, B. Parker, R. (2001a) Simplified scoring and psychometrics of the revised 12-item Short-Form Health Survey. *Outcomes Management for Nursing Practice*. 5(4), p161-166.

Resnick, B. Nahm, E. S. (2001b) Reliability and validity testing of the revised 12-item Short-Form Health Survey in older adults. *Journal of Nursing Measurement*. 9(2), p151-161

Sanchez-Arenas, R., Vargas-Alarcon, G., Sanchez-Garcia, S., Garcia-Peña, C., Gutierrez-Gutierrez, L. et. al. (2014). Value of EQ-5D in Mexican city older population with and without dementia (SADEM study). *International Journal of Geriatric Psychiatry*. 29(5), p478-488.

Shou, J., Ren, L., Wang, H., Yan, F., Cao, X. et. al. (2016) Reliability and validity of 12-item Short-Form health survey (SF-12) for the health status of Chinese community elderly population in Xujiahui district of Shanghai. *Aging Clinical & Experimental Research*. 28(2) p339-346.

Tidermark, J., Zethraeus, N., Svensson, O., Törnkvist, H., Ponzer, S. (2002). Femoral neck fractures in the elderly: functional outcome and quality of life according to EuroQol. *Quality of life research*. 11(5), p473-481.

Tidermark, J., Bergström, G., Svensson, O., Törnkvist, H., Ponzer, S. (2003). Responsiveness of the EuroQol (EQ-5D) and SF- 36 in elderly patients with displaced femoral neck fractures. *Quality of life research*. 12(8), p1069-1079.

Tidermark, J., Bergström, G. (2007). Responsiveness of the EuroQol (EQ-5D) and the Nottingham Health Profile (NHP) in elderly patients with femoral neck fractures. *Quality of life research*. 16(2), p321-330.

van Leeuwen, K. M., Bosmans, J. E., Jansen, A., Rand, S. E., Towers, A. et. al. (2015a). Dutch translation and cross-cultural validation of the Adult Social Care Outcomes Toolkit (ASCOT). *Health and Quality of Life Outcomes*. 1, p56

van Leeuwen, K. M., Jansen, A., Muntinga, M. E., Bosmans, J. E., Westerman, M. J. (2015b). Exploration of the content validity and feasibility of the EQ-5D-3L ICECAP-O and ASCOT in Older adults. *BMC Health Services Research*. 15, p201.

van Leeuwen, K. M., Bosmans, J. E., Jansen, A. P. D., Hoogendijk, E. O., van Tulder, M. W. et al. (2015c). Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. *Value in Health*. 18(1), p35-43.

Appendix 6 – Characteristics of included studies

Reference	Country	Population/ Setting/ Study/ Intervention	Administration	PROM	Language	n	Mean age (SD)[range]	Gender % female
Bentur et al. 2010	Israel	community dwelling older people (70+)	face to face interview	SF-12	Hebrew	399	78.2	53
Brazier et al. 1996	UK	women (75+) recruited from a double blind prospective RCT of clodronate	interview self-complete	EQ-5D-3L	English	370 (123 follow up)	80.1 (4.5)	100
Cernin et al. 2010	USA	community dwelling older (60+) African American adults	telephone	SF-12 think v2	English	985	71.0 (7.3)	71.6
Coast et al. 1998	UK	acute care patients (65+) acting as participants in RCT comparing hospital at home and routine hospital care	either self or interviewer administrated	EQ-5D-3L	English	214	median 79 IRQ 74-84	70
Davis et al. 2012	Canada	community dwelling older adults (70+) with mobility impairments visiting the Vancouver Falls Prevention Clinic	interview administered (not clear who filled out)	EQ-5D-3L	English	215	78.7 (6.2)	71.6
Diaz-Redondo et al. 2014	Spain	institutionalised older adults (60+) with dementia	completed by care giver	EQ-5D-3L	Spanish	525	85.6 (6.7)	83
Fleishman et al. 2003	USA	non-institutionalised adult respondents from 2000 Medical Expenditure Panel Survey	self-complete	SF-12	English	11626	age groups and proportions	55
Hackert et al. 2017	UK	Older social care users (70+)	online	ASCOT	English	205	76.0 (5.5)	50

Reference	Country	Population/ Setting/ Study/ Intervention	Administration	PROM	Language	n	Mean age (SD)[range]	Gender % female
Holland et al. 2004	UK	older people (79+) admitted as an emergency and taking 2+meds a day, participating in RCT of home-based medication review	baseline - self-complete in interview assisted by recruiter. Follow up post	EQ-5D-3L	English	145	84.7	57
Jakobsson et al. 2007, Jakobsson et al. 2012	Sweden	general older people (75+) (including those in community and in special accommodation e.g. nursing homes) and stroke patients	general group = postal questionnaire. stroke = interview	SF-12	Swedish	general = 4278, stroke = 89	general =83.7 (5.7) [75-105], stroke = 77.2 (6.7)	general=61.1 stroke=49.4
Kaambwa et al. 2015	Australia	community dwelling older people (65+), cognitively intact, receiving aged care services	self-completion at a group interview	EQ-5D-3L ASCOT	Australian	87	80 [range 65-93]	66
Kunz et al. 2010	Germany	mild-moderate dementia patients cared for in the family home, participating in a cluster-randomised trial of whether further training of GPs and the offer of family counselling can delay institutionalisation	patients - interview carers - computer aided telephone interview	EQ-5D-3L	German	patients = 390(carers = 357)	80.2 (6.7)	67.5
Liang et al. 2014	China	empty nested elderly (60+) rural China	one-to-one interview - interviewer admin	EQ-5D-3L SF-12	Chinese version	967	78.3 (9.6)	45.7
Lung et al. 2017	Australia	older people (65+) living permanently in nursing homes in an RCT of nurse led care coordination in nursing home	interview	EQ-5D-3L	English (Aus. weights)	199	85.1 (8.9)	75.4

Reference	Country	Population/ Setting/ Study/ Intervention	Administration	PROM	Language	n	Mean age (SD)[range]	Gender % female
Lutomski et al. 2017	Netherlands	community dwelling older people (65+) - four geriatric conditions: hearing conditions, joint damage, urinary incontinence, dizziness with falls and a healthy group. Data from TOPICS-MDS dataset - public access data repository on health and wellbeing of older persons and informal carers	not clear - a large dataset of pooled information from many studies	EQ-5D-3L	Dutch	25637	78 (6)	58.3
Malley et al. 2012	UK	older people (65+) receiving publicly funded home care services	face to face computer assisted interview home	ASCOT	English	301	age groups and proportions given	68
Netten et al. 2012	UK	older people (65+) using publicly funded home care services. Recruited through user experience survey, interview face to face computer assisted	face to face computer assisted interview	ASCOT	English	301	age groups and proportions given	68
Parsons et al. 2014	UK	older people (60+) with hip fracture from Warwick hip trauma study RCT	baseline interview follow-up telephone	EQ-5D-3L	English	whit baseline = 151 white baseline = 236	WHIT 83.1 WHITE 83.6	WHIT 71 WHITE 75
Pettit et al. 2012	UK	community dwelling older people (65+)	Home interview	SF-12	English	541	median =74 range 65-102	58

Reference	Country	Population/ Setting/ Study/ Intervention	Administration	PROM	Language	n	Mean age (SD)[range]	Gender % female
Ratcliffe et al. 2017	Australia	post-hospital population of frail older people living in residential aged care. From an RCT of multidisciplinary rehab services to hip fracture patients	interview	EQ-5D-5L	English	240	88.6 (5.6)	74.2
Resnick et al. 2001	USA	study 1 older adults in continuing care retirement community (65+) and study 2 older adults (65+) discharged from an acute care setting	face-to-face interview in retirement community and telephone in acute care discharge	SF-12	English	Retirement community = 187. Acute discharge = 211	Retirement community = 86. Acute discharge = 73	Retirement community = 78. Acute discharge = 60
Resnick et al. 2001b	USA	older adults in continuing care retirement community (65+)	face-to-face interview in retirement community	SF-12	English	Retirement community = 185	86 (6.1)	82
Sanchez-Arenas et al. 2014	Mexico	community dwelling older adults (60+)	face-to-face interview	EQ-5D-3L	Mexico	normal cog = 2796 dementia = 109	normal cog = 71.0 dementia = 78.5	normal cog = 57.4 dementia = 64.2
Shou et al. 2016	China	community dwelling older (65+)	face-to-face interview	SF-12	Chinese version	1343	72 (7.36) [65-80]	57.2
Tidermark et al. 2002	Sweden	older people (65+) with femoral neck fractures living independently	face-to-face interview	EQ-5D-3L	Swedish	67 followed up	79.9 (7.3)	76
Tidermark et al. 2003	Sweden	older people with displaced femoral neck fractures living independently	face-to-face interview	EQ-5D-3L	Swedish	95 followed up	approximately 80	81

Reference	Country	Population/ Setting/ Study/ Intervention	Administration	PROM	Language	n	Mean age (SD)[range]	Gender % female
Tidermark et al. 2007	Sweden	older women (70+) with femoral neck fractures living independently.	face-to-face interview	EQ-5D-3L	Swedish	60	83 (5)	100
Van Leeuwen et al. 2015a, Van Leeuwen et al. 2015b, Van Leeuwen et al. 2015c	Netherlands	Subset of patients from ACT study - stepped wedged cluster RCT evaluation of a geriatric care model for frail older adults (65+) living at home.	content validity - think aloud interviews. Other properties home computer assisted interviews	ASCOT	Dutch	content validity =10 other=190	content val=[75-100] other=82.4	content val=60 other=71.6

Appendix 7 – Methodological quality of included studies

Study	Country	CTT or IRT	Validity				Reliability		Responsiveness
			Content	Structural	Construct	Cross-cultural	Internal Consistency	Test-retest/Inter rater	
EQ-5D-3L									
Brazier et al. 1996	UK	CTT	-	-	Good	-	-	Fair	Fair
Coast et al. 1998	UK	CTT	-	-	Good	-	-	-	Fair
Davis et al. 2012	Canada	CTT	-	-	Fair	-	-	-	-
Diaz-Redondo et al. 2014	Spain	CTT and Rasch	-	Fair	Good	-	Fair	Fair	-
Holland et al. 2004	UK	CTT	-	-	Fair	-	-	-	Fair
kaambwa et al. 2015	Australia	CTT	-	-	Good	-	-	-	-
Kunz et al. 2010	Germany	CTT	-	-	Good	-	-	Fair	Good
Liang et al. 2014	China	CTT	-	-		-	Fair	-	-
Lung et al. 2017	Australia	CTT	-	-		-	-	-	Fair
Iutomski et al. 2017	Netherlands	CTT	-	-	Good	-	-	-	-
Parsons et al. 2014	UK	CTT	-	-	Fair	-	-	-	Good
Sanchez-Arenas et al. 2014	Mexico	CTT	-	-	Poor	-	Fair	-	-
Tidermark et al. 2002	Sweden	CTT	-	-	Fair	-	-	-	-
Tidermark et al. 2003	Sweden	CTT	-	-	-	-	-	-	Good
Tidermark et al. 2007	Sweden	CTT	-	-	-	-	-	-	Good
Van Leeuwen et al. 2015a and Van Leeuwen et al. 2015c	Netherlands	CTT and qualitative	Excellent	-	Good	-	-	Good	-
EQ-5D-5L									
Ratcliffe et al. 2017	Australia	CTT	-	-	Good	-	-	-	-
SF-12									
Bentur et al. 2010	Israel	CTT	-	Poor	Fair	-	Fair	-	-
Cernin et al. 2010	USA	CTT	-	Good	Good	-	Good	-	-

Study	Country	CTT or IRT	Content	Structural	Construct	Cross-cultural	Internal Consistency	Test-retest/Inter rater	Responsiveness
Fleishman et al. 2003	USA	IRT/DIF	-	Excellent	-	-	-	-	-
Jakobsson et al. 2007 and 2012	Sweden	CTT	-	Poor	Fair	-	Good	-	-
Liang et al. 2014	China	CTT	-	-	-	-	Good	-	-
Pettit et al. 2012	UK	CTT	-	-	Good	-	-	-	-
Resnick et al. 2001	USA	CTT	-	Fair	Good	-	Good	Fair	-
Resnick et al. 2001b	USA	CTT	-	Good	Good	-	Good	-	-
Shou et al. 2016	China	CTT	-	Good	Fair	-	Poor	-	-
ASCOT									
Hackert et al. 2017	UK	CTT	-	Poor	Fair	-	-	-	-
kaambwa et al. 2015	Australia	CTT	-	-	Good	-	-	-	-
Malley et al. 2012 and Netten et al. 2012	UK	CTT	-	-	Good	-	-	-	-
Van Leeuwen et al. 2015a and Van Leeuwen et al. 2015b Van Leeuwen et al. 2015c	Netherlands	CTT and qualitative	Excellent	-	Good	Good	-	Good	Good

Appendix 8– Included study results

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
EQ-5D-3L								
Brazier et al. 1996 (UK)			Known group - Sig diff in index scores: recent GP, inpatient, LTCs. Insig diff in index scores: age, recent outpatient, A&E, OPCS severity category		Test Retest (3month) index corr=0.67 VAS corr=0.53. Mean diff (95% CI) index=0.1 (-2.81, 2.60) VAS=-4.97 (-9.74, -0.20)	Hypothesised improvement in health from recent service use to not ES=0.23-0.42 (small). hypothesised improvement in health from having a LTC to not ES=0.85 (Strong)	<10% missing per item	No floors/ceilings on overall score - no details of item distributions
Coast et al. 1998 (UK)			Known group - Sig diff in VAS scores: age and LTC. No Sig diff in index scores. Convergent - 4/10 expected relationships between EQ-5D domains and Barthel at baseline - many more sig at 4-weeks. Sig relationships with COOP WONCA			As expected, mean scores showed most improvement in least severe conditions expected to recover more quickly (elective knee surgery) and least improvement in most severe conditions (stroke). No sig tests as numbers small.	<5% missing per item. Higher for VAS	Ceiling for anxiety/ depression (66%). Floor for usual activities (47%). Worst category for self-care (4%) and anxiety/depression (5%)

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Davis et al. 2012 (Canada)			Convergent - EQ-5D index corr with ICECAP=0.47 sig moderate. No sig corrs between EQ-5D domains and PPA or MMSE. Mobility sig corr with SPPB and Self Care sig corr with IADLs but both corrs<0.25 (weak)				Not mentioned	Not mentioned
Diaz- Redondo et al. 2014 (Spain)		EFA PCA (varimax) 2 factors account for 67% variance - F1 (functional) - mob, SC UA loadings 0.74, 0.85, 0.82. F2 (subjective) - Pain Anx 0.78, 0.82. Lack of unidimensionality confirmed by Rasch lack of fit	Known group - Sig diff in index scores: gender, age, functional status, comorbidities, depression (CDR score). Insig diff in index scores: education. Convergent - EQ-5D index and VAS sig corr with QOL-AD and QUALID - 0.38<corr<0.58 (Mod-strong)	$\alpha=0.64$ (good>0.7). 0.21<ITCC< 0.53	Inter-rater ICC(with carer response)=0.72 (good>0.7)		<3% missing per item	No score floor/ceilings. But big item ceilings for pain and anxiety/ depression. Floors for self-care and usual activities
Holland et al. 2004 (UK)			Known group - Sig diff in index scores: gender, number of drugs prescribed at discharge. Insig diff in index score: age, social class, living alone.			ES(0-6 month)=0.55 (moderate)	Completion rate 81%	Worst categories hardly used for mobility (2%) and anxiety/ depression (3%)

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
kaambwa et al. 2015 (Australia)			Known group - Sig diffs in index scores: age(unexpected direction), gender, levels general health. Insig diffs in index scores: living arrangement, education, informal care. Convergent - EQ-5D index corr with ASCOT score=0.50 and corr with OPQOL score=0.53 (moderate). Weak corrs between EQ-5D and ASCOT domains (biggest corr=0.35). Some mod corrs between EQ-5D and OPQOL domains				0% missing	Score ceiling of 15%. Item ceilings for self-care and anxiety/ depression (>50%). Worst category hardly used for mobility (2%), self-care (3%) and anxiety/ depression (2%)
Kunz et al. 2010 (Germany)			Convergent - EQ-5D index (patient report) sig corr Barthel=0.50 (mod), IADL=-0.4 (mod) and MMSE=0.18 (weak). EQ-5D index (proxy report) sig corr Barthel=0.67 (strong), IADL=-0.57 (strong) and MMSE=0.24 (weak).		Inter-rater - ICC=0.48. ICC(mild dementia)=0.54. ICC(mod dementia)=0.33. (good>0.7) Mean diff in scores=0.1 (sig - carers proxy score lower than patient self-rated)	ES(0-1yr in patients whose GP reported health increased) = 0.12 (small). ES(0-1yr in patients whose GP reported health decreased)=0.41 (small)	Approximately 3% missing per item	Item ceilings for mobility, self-care and anxiety/ depression
Liang et al. 2014 (China)				EQ-5D whole $\alpha=0.775$ (good>0.7)			Not mentioned	Item floors for all items (>47%)

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Lung et al. 2017 (Australia)						ES(0-6 month)= 0.19 (small)	Only analysed fully complete - no further details	Not mentioned
Iutomski et al. 2017 (Netherlan ds)			Known group - Sig diffs in index scores: age, gender, education, comorbidities. Insig diffs in index scores: marital status, living alone. Convergent - Mobility mod corr Katz waking items. Self care strong corr Katz bathing/dressing items. Usual Acts strong corr Katz IADL summary score. Anx mod-strong corr mental health summary SF-36. EQ-5D index weak-mod corr Cantrils Ladder QoL score.				Not mentioned	Score ceiling of 19%. Big item ceilings for self-care, usual activities and anxiety/ depression (>56%)
Parsons et al. 2014 (UK)			Convergent - EQ-5D index corr OHS=0.74 (strong) and ICECAP=0.34 (weak-mod)			Study 1 ES(0-6wks)=0.68 (moderate). ES(0-12wks)=0.32 (small). ES(0-52wks)=0.27 (small). Study 2 ES(0-4wks)=0.64 (moderate). ES(0-4months)=0.3 (small).	Not mentioned	Not mentioned

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Sanchez- Arenas et al. 2014 (Mexico)			Known group - Insig diff in index scores: with/without dementia. Convergent - index corrs mod-strong with SF-36 domains (0.43-0.79), strong with ADLs and IADLs (0.60-0.71), weak-mod with Charlsons index (0.26-0.36) and weak for MMSE (0.07-0.14) in both normal capacity and dementia groups	normal capacity sample $0.71 < \alpha < 0.83$. Dementia sample $0.72 < \alpha < 0.83$ (good>0.7)			Not mentioned	Gen pop group - big item ceilings for mobility, self-care, usual activities and anxiety/depression. Worst category hardly used for mobility (1%), self-care (1%), usual activities (3%) and anxiety/depression (4%). Dementia group - ceiling effect for self-care
Tidermark et al. 2002 (Sweden)			Known group - sig diffs in EQ-5D scores according to cut-offs for pain, mobility, ADL limitations, living status (community or not - sig at 4 month only not 17month)				Not mentioned	Not mentioned

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Tidermark et al. 2003 (Sweden)						Responsiveness - Mean EQ-5D at 4 months - IF group=0.73 THR group=0.60 (sig diff). According to EC 53% patients good early clinical outcome and 47% less good. 71% THR group and 34% IF group good early outcome (sig). Sig diff in EQ-5D scores at 4 months in those with good and less good outcome. EQ-5D change those with less good outcome at 4 months =-0.26 (SD 0.29) SES=1.37 (large) SRM=0.9 (large). Corr between EQ-5D and SF-36 change scores=0.39	98.9% completion rate	Not mentioned

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Tidermark et al. 2007 (Sweden)						Responsiveness - change scores (prefracture-6months) among displaced fracture patients high and sig - SRM large=1.14. No sig change in scores of undisplaced fractures. Possible to use change scores to discriminate between displaced and undisplaced fractures - 74.5% correctly classified and using logistic regression the risk of having a displaced fracture increases as the change score increases (-ve)	98% completion rate	Score ceiling 22%
Van Leeuwen et al. 2015a and Van leeuwen et al. 2015c (Netherlan ds)	EQ-5D - Pain/discomfort, anxiety/depression, usual activities often interpreted too narrowly. Wanted more options for mobility (3L). Often perceived positive answering where the interviewer felt given their knowledge of the participant they		Convergent - Strong corr with SF-12 PCS (0.60), moderate corrs with ICECAP-O, ASCOT, Global Health rating scale, QOL rating scale, ADLs, SF-12 MCS, Pearlin Mastery scale (0.34-0.5) and trivial corr with Client Centredness Questionnaire (0.02).		Test-retest (1-2 weeks) - ICC(95%CI) =0.79 (0.72, 0.85)	Responsiveness - Correlations between change scores (12months-18months) with other measures weak (0.01-0.23) - strongest corr in change scores with SF-12 PCS (0.23)	0% EQ-5D scores missing	Not mentioned

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
	had selected a more positive answer. Respondents found questions relevant to their QoL and easy to answer							
EQ-5D-5L								
Ratcliffe et al. 2017 (Australia)			Convergent = Sig weak corr with MMSE (cognition) score (expected), weak sig corr with CDSS (depression) and moderate sig corrs with MBI (functioning) and PainAd (pain). All insig at 4 weeks. Known group = EQ-5D scores at baseline across CSDD (depression), MBI (functioning) and PainAd (pain) thresholds small to moderate effect sizes (<0.3) (expected). Insig relationship MMSE. All insig at 4 weeks.				Not clear (many rated by proxy)	Not mentioned
SF-12								
Bentur et al. 2010 (Israel)		Found 3 factors explaining 68% to variance (F1=physical functioning, pain, general health, vitality, social funct) (F2=emot role and mental health)	Convergent - MCS strong corr with GDS (-0.67). PCS strong corr with Barthel (0.61) and IADL (-0.68).	Whole SF-12 $\alpha=0.89$. Found 3 factors inconsistent with SF-12 structure - Factor 1 $\alpha=0.86$. F2			Not mentioned	Ceilings for physical role domain, emotional role domain and social functioning domain -

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
		(F3=physical role questions). No fit statistics provided		$\alpha=0.71$. F3 $\alpha=0.85$				version not clear
Cernin et al. 2010 (USA)		Two factors present. F1 - general health, physical functioning, physical role, pain, vitality, social functioning. F2 - emotional role, mental health, social functioning. Social functioning loaded on both. No fit statistics provided	Known group - Sig diffs in both PCS and MCS scores: long-term conditions, number of prescription meds, recent GP visit, recent hosp inpatient, activity level, need for home care, enjoyment of senior life. Sig diffs in MCS score: nursing home days. Insig: recent A&E admission.	Whole SF-12 $\alpha=0.77$ PCS (Q1-5,8) $\alpha=0.45$. MCS (Q6,7, 9-12) $\alpha=0.76$.			Not mentioned	Not mentioned
Fleishman et al. 2003 (USA)		Differential item functioning - Older people tend to rate themselves more highly on calm/peaceful and energy (and downhearted and social acts) and lower on moderate activities and stairs than would be expected from their underlying physical and mental health (direct DiF effects on items)					14% missed at least one item	Not mentioned

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Jakobsson et al. 2007 Jakobsson et al. 2012 (Sweden)		General older EFA= 3 factors (F1=Q1-5,8) (F2= Q9-12) (F3=6,7). When forced a 2 factor solution items 10 and 12 had high loadings on both factors. Stroke EFA 3 factors - (F1=Q2-5) (F2=Q6,7,9,11,12) (F3=1,8,10). When forced 2 factors Q10 higher on phys and Q1 high on both. Goodness of fit outside acceptable ranges.	Convergent - MCS moderate corr with Nervous/worry (-0.44) and Depressed mood (-0.49). PCS strong corr with IADL (-0.58) and Walking problems (-0.61) and moderate corr with PADL (-0.43) and Pain (0.44)	General older - PCS $\alpha=0.85$. MCS $\alpha=0.76$ (MCS $\alpha=0.83$ in 2012 paper report). Stroke - PCS $\alpha=0.82$. MCS $\alpha=0.78$			14% missed at least one item	Items with highest rate of missing - emotional role carefully (7.9%) and emotional role limited (5.9%)
Liang et al. 2014 (China)				PCS $\alpha=0.71$. MCS $\alpha=0.76$.			Not mentioned	Floors all domains

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Pettit et al. 2012 (UK)			Known group - Sig diffs in both PCS and MCS scores: self-reported health problems, ADL limitations, receiving services, impaired vision and depression. Sig diffs PCS only: impaired hearing and dementia. Sig diffs MCS only: self-reported psychiatric problems. Convergent - (forward linear regression methods) MCS accounted for more variation in depression subscale than PCS (-0.65 vs -0.12). PCS accounted for more variation in ADL limitation scale with MCS (-0.72 vs -0.16). Neither performed well in dementia.				94.5% completion rate	Not mentioned
Resnick et al. 2001 (USA)		Run preassumed 2 factor CFA, assuming item 10 loads on physical and item 12 on both. For general population sample low item 12 loadings (PCS =0.34 and MCS=0.41) and discharge sample item 12 loading on MCS=0.19 very low - wouldn't say this item loads onto	known group - Sig diffs in both PCS and MCS scores (revised structure): number of long-term conditions. Sig PCS only: regular exercise	General population group PCS $\alpha=0.87$ MCS $\alpha=0.8$. Recent acute hospital group - PCS $\alpha=0.81$ and MCS $\alpha=0.72$ (PCS and MCS using revised structure)	Test retest - general population group (2-4 weeks). Sig corr PCS scores=0.86 and sig corr MCS scores=0.73 (revised structure)		Not mentioned	Not mentioned

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
		MCS. RMSEA=0.09 both models						
Resnick et al. 2001b (USA)		CFA= vitality wanted to load into physical factor and social functioning wanted to load onto both - fit still poor RMSEA=0.14 but improved on the initial split with RMSEA=0.17	Known group - Sig diff in both PSC and MCS (revised structure) scores: number of long-term conditions and regular exercise	Original split of items PCS $\alpha=0.84$ and MSC $\alpha=0.7$. Revised item split physical factor $\alpha=0.89$ mental factor $\alpha=0.7$.				
Shou et al. 2016 (China)		2 Factors (F1 Q1,2,3,4,5,6,7,8,12) (F2 Q8,9,10,11,12) pain and social load both (both higher on phys). RMSEA=0.041	known group - Sig diffs in both PCS and MCS scores: age, education level, economic status and long-term conditions. Insig: marital status and gender.	$\alpha=0.91$ whole SF-12			97.7% completion rate	Not mentioned
ASCOT								
Hackett et al. 2017 (UK)		Together ASCOT and ICECAP-O items led to a 3-factor model (F1=Acontrol, Aoccupation, Asocial part, Isecurity, Irole, Ienjoyment, Icontrol = 27% explained var)(F2=Apersonal, Afood 10% explained var) (F3=Aaccom,	Known group - on average ASCOT score higher in those with above average health (EQ-5D-5L, GDS-15, Barthel) and wellbeing (OPQOL-13, SWLS, Cantrills Ladder). ASCOT score also increasing with age. Not clear if no significance test or if failed a significance test. Convergent - strong corrs with EQ-5D-5L, GDS-15, OPQOL-13, SWLS, Cantrills Ladder (>0.6) and				0% missing - online survey so had to answer fully before could move on	Ceilings for all items (at least 35%). Worst level for each item hardly used (<4%)

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
		I attachment 10% explained var). Asafety and Adignity didn't load on any factor higher than 0.4. Cannot be sure what factor structure would be if measures separated	moderate corr with Barthel Index (0.45)					
kaambwa et al. 2015 (Australia)			Known group - Sig increase in ASCOT score in those with better self-reported general health. Insig increase in ASCOT score with age and men scores higher. All others no clear relationship (living situation, education, income, informal care) Convergent - Moderate-strong corrs with EQ-5D-5L (0.5) and OPOL-Brief (0.58). Correlations between relevant dimensions low-moderate rather than strong as hypothesised				0% missing	6% received top score on ASCOT. Big item ceilings for personal cleanliness/comfort, food/drink, safety, accommodation and dignity. All worst categories hardly used (2% or less)

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Malley et al. 2012 and Nettten et al. 2012 (UK)			<p>Convergent - Sig relationships found between ASCOT items and GHQ-12, overall QoL, CASP control and autonomy subscale, EQ-5D-3L and UCLA loneliness scale.</p> <p>Dignity the only question where some relationships weren't sig. Other variables such as demographics, disability, environment/locality, social contact and support, participation and service quality were also mostly found to have the expected relationships with ASCOT items. Items with weakest evidence of validity were food/drink and dignity.</p> <p>Food/drink wording changed after analysis (and accommodation wording). Hard to find comparators for dignity</p>				item level missing rates ranged from 10.3% for control to 9.3% for personal cleanliness/comfort, safety and dignity.	Large ceilings personal cleanliness/comfort, food/drink, safety, accommodation, social participation and dignity. Worst categories used by <5% for all items except social participation

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
Van Leeuwen et al. 2015a, Van Leeuwen et al. 2015b, Van Leeuwen et al. 2015c (Netherlands)	items generally understood. ASCOT safety and accommodation answered too narrowly. "safety" focus on crime - changed to "feel safe and secure". some issues in understanding/ interpretation of occupation and control - changed to "able to arrange your daily life". dignity question confusing - most didn't see how support and care would influence the way they think about themselves. often skipped dignity questions (10% missing rather than 2%) - changed to "having help affects my self- image and the way I'm helped makes me feel respected." some difficulties in		Convergent - Strong corr with ICECAP-O (0.63), mod corrs with EQ-5D-3L, Global Health rating scale, QOL rating scale, ADLs, SF-12 MCS, Pearlin Mastery scale (0.34-0.5) and weak corrs with SF-12 PCS and Client Centredness Questionnaire (0.26 and 0.22).		Test-retest (1-2 weeks) - ICC(95%CI) =0.71 (0.60, 0.78)	Responsiveness - Correlations between change scores (12months-18months) with other measures weak (0.02-0.34) - strongest corr in change scores with Client Centredness Questionnaire (dignity) (0.34)	14.7% ASCOT index scores missing - mostly due to dignity item (missing 12.6%)	Ceilings all items except maybe social (29%). Worst level hardly used

Study (country)	Content Validity	Structural Validity	Construct Validity	Internal Consistency Reliability	Test-retest/ Inter rater Reliability	Responsiveness	Missing data rates	Response Distributions
	seeing differences between response options (first 2 options in occupation) and misunderstanding of the best food/drink option. issues with response options having multiple sections, which only half fitted the respondent e.g. last option in social. Positive responding seen							

Sig=Significant Insig=insignificant LTC=Long-term condition Corr=correlation ES=effect size Mob=mobility SC=self care UA=usual activities Anx=anxiety/depression

Mod=moderate ICC=Intraclass correlation EC=external criterion SRM=standardised response mean

Chapter 4 appendices

Appendix 9 – Original response distributions of ONS-4

response option	life satisfaction		worthwhile		happy		anxious	
	n	%	n	%	n	%	n	%
0	205	3.2	142	2.2	153	2.4	2,453	38.6
1	106	1.7	105	1.7	111	1.7	643	10.1
2	216	3.4	201	3.2	209	3.3	612	9.6
3	336	5.3	284	4.5	265	4.2	490	7.7
4	364	5.7	279	4.4	335	5.3	371	5.8
5	710	11.2	555	8.7	575	9.1	554	8.7
6	535	8.4	435	6.8	471	7.4	283	4.5
7	902	14.2	732	11.5	805	12.7	320	5.0
8	1,318	20.8	1,299	20.5	1,258	19.8	264	4.2
9	925	14.6	1,084	17.1	1,124	17.7	117	1.8
10	628	9.9	1,104	17.4	944	14.9	141	2.2
missing	106	1.7	131	2.1	101	1.6	103	1.6

These are the original response distributions for the ONS-4 of the whole HIPO sample. Categories in bold were chosen for merging as they represent the least used pair of adjacent categories and they all represent the lowest level of wellbeing for that question (a response of 0 on the ONS-4 questionnaire signals not at all anxious while 10 indicates completely anxious). The anxiety question was then reverse coded so that it matched the other questions in that higher numbered responses indicate higher levels of wellbeing.

Appendix 10 – ScHARR Ethics approval letter



Downloaded: 16/11/2018
Approved: 05/04/2017

Hannah Penton
Registration number: 150110521
School of Health and Related Research
Programme: PhD

Dear Hannah

PROJECT TITLE: Secondary data analysis - Validating commonly used QoL measures in elderly populations
APPLICATION: Reference Number 012758

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 05/04/2017 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 012758 (dated 17/03/2017).

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

Appendix 11 – HIPO development and validation sample characteristics

Characteristic n (%)	Development	Validation	P-value
N	3176	3175	
Gender			0.106
Female (0)	1626 (51.2)	1561 (49.2)	
Age Group			0.872
>25	98 (3.1)	95 (3.0)	
25-34	201 (6.3)	176 (5.5)	
35-44	273 (8.6)	268 (8.4)	
45-54	491 (15.5)	516 (16.3)	
55-64	765 (24.1)	749 (23.6)	
65-74	677 (21.3)	701 (22.1)	
75-84	495 (15.6)	501 (15.8)	
85+	176 (5.5)	169 (5.3)	
Age Group			0.552
18-64	1828 (57.6)	1804 (56.8)	
65+	1348 (42.4)	1371 (43.2)	
Marital Status			0.381
Married	1818 (58.8)	1802 (58.4)	
Cohabiting	190 (6.2)	208 (6.7)	
Single	435 (14.1)	395 (12.8)	
Divorced/Separated	293 (9.5)	302 (9.8)	
Civil Partnership	13 (0.4)	21 (0.7)	
Widowed	341 (11.0)	359 (11.6)	
Employment Status			0.345
Full-time	781 (26.9)	775 (26.6)	
Part-time	331 (11.4)	318 (10.9)	
Unemployed/seeking work	68 (2.3)	56 (1.9)	
Housework	133 (4.6)	104 (3.6)	
Student	45 (1.6)	43 (1.5)	
Retired	1216 (41.9)	1281 (44.0)	
Long-term Sick	328 (11.3)	332 (11.4)	
General Health			0.996
Poor	366 (11.7)	363 (11.6)	
Fair	792 (25.3)	798 (25.5)	
Good	954 (30.5)	947 (30.3)	
Very Good	798 (25.5)	795 (25.4)	
Excellent	220 (7.0)	227 (7.3)	
EQ-5D-5L Utility	0.67 (SD 0.28)	0.67 (SD 0.29)	0.669
EQ-5D VAS	69.3 (SD 22.2)	69.0 (SD 23.1)	0.874
SF-6D Utility	0.69 (SD 0.16)	0.69 (SD 0.16)	0.987
Wellbeing VAS	69.3 (SD 23.9)	69.1 (SD 24.5)	0.982
Accommodation			
Community	3007 (94.7)	3018 (95.1)	
Nursing/residential home	44 (1.4)	33 (1.1)	0.207
Informal Care			
Received>0 (%)	796	790	0.993
Mean Hours last week*	28.6 (SD 44.3)	33.4 (SD 47.7)	0.510

Characteristic n (%)	Development	Validation	P-value
Formal Care			
Received>0 (%)	396	360	0.370
Mean Hours last week*	11.2 (SD 30.3)	12.0 (SD 29.5)	0.437

Appendix 12 – HSE development and validation sample characteristics

Characteristic n (%)	Development	Validation	P-value
N	3659	3594	
Gender			0.402
Female	2034 (55.6)	2033 (56.6)	
Age Group			0.708
>25	271 (7.4)	249 (6.9)	
25-34	519 (14.2)	508 (14.1)	
35-44	674 (18.4)	635 (17.7)	
45-54	693 (18.9)	691 (19.2)	
55-64	562 (15.4)	576 (16.0)	
65-74	579 (15.8)	541 (15.1)	
75-84	276 (7.5)	298 (8.3)	
85+	85 (2.3)	96 (2.7)	
Age Group			0.752
18-64	2719 (74.3)	2659 (74.0)	
65+	940 (25.7)	935 (26.0)	
Marital Status			0.651
Married	1962 (53.6)	1943 (54.1)	
Cohabiting	427 (11.7)	406 (11.3)	
Single	641 (17.5)	608 (16.9)	
Divorced	275 (7.5)	266 (7.4)	
Separated	91 (2.5)	79 (2.2)	
Civil Partnership	5 (0.1)	2 (0.1)	
Widowed	257 (7.0)	289 (8.0)	
Employment Status			0.216
Employed (employee)	1768 (48.3)	1689 (47.0)	
Employed (self-employed)	295 (8.1)	293 (8.2)	
Unemployed	111 (3.0)	134 (3.7)	
Retired	959 (26.2)	995 (27.7)	
Other economically inactive	524 (14.3)	480 (13.4)	
General Health			0.364
Very bad	61 (1.7)	69 (1.9)	
Bad	197 (5.4)	199 (5.5)	
Fair	671 (18.3)	626 (17.4)	
Good	1540 (42.1)	1553 (43.2)	
Very good	1190 (32.5)	1144 (31.8)	
EQ-5D-5L Utility	0.86 (SD 0.23)	0.85 (SD 0.23)	0.827
EQ-5D VAS	77.8 (SD 18.2)	77.4 (SD 18.8)	0.771
Informal Care			
Received>0 (%)	145	169	0.122
Mean Hours last week*	20.7 (SD 30.5)	16.6 (SD 26.2)	0.304
Formal Care			
Received>0 (%)	37	41	0.593
Mean Hours last week*	10.6 (SD 17.2)	7.1 (SD 9.3)	0.521

Appendix 13 – ASCS development and validation sample characteristics

Characteristic n (%)	Development	Validation	P-value
N	34537	34481	
Gender			0.790
Female	20688 (59.9)	20691 (60.0)	
Age Group			0.272
18-64	13709 (39.7)	13547 (39.3)	
65+	20823 (60.3)	20932 (60.7)	
General Health			0.946
Very bad	1508 (4.5)	1533 (4.6)	
Bad	4458 (13.2)	4417 (13.1)	
Fair	13605 (40.3)	13492 (40.1)	
Good	9243 (27.4)	9225 (27.4)	
Very good	4950 (14.7)	4977 (14.8)	
Global QoL			0.910
Very bad	1042 (3.1)	1052 (3.1)	
Bad	1933 (5.7)	1948 (5.8)	
Fair	9672 (28.5)	9690 (28.7)	
Good	10758 (31.7)	10782 (31.9)	
Very good	10486 (30.9)	10347 (30.6)	
Support Setting			0.787
Community	26112 (75.6)	26046 (75.5)	
Residential Home	6576 (19.0)	6551 (19)	
Nursing Home	1844 (5.3)	1882 (5.5)	
Informal Care			
Received>0 (%)	33115	33112	0.252
Formal Care			
Received>0 (%)	34537 (100)	34481 (100)	

Appendix 14 – EQ-5D IRT MPlus model input files for each stage of analysis plus final model for other measures

Appendix 15 – Model fit statistics for each model tested in the DiF identification process

EQ-5D-5L

Model	Parameter	Free	Chi ²	Chi ² DF	p-value	CFI	RMSEA	RMSEA Higher CI	RMSEA Lower CI	DIFFTEST p-value
Configural		50	112.8	10	<0.000	0.998	0.08	0.068	0.095	<0.000
Non-uniform A		46	93.2	14	<0.000	0.998	0.06	0.049	0.072	0.001
Non-uniform B (pain)		47	88.6	13	<0.000	0.998	0.061	0.049	0.073	0.008
Non-uniform C (self-care)		48	76.1	12	<0.000	0.999	0.058	0.046	0.071	0.687
Uniform A		29	425.8	31	<0.000	0.992	0.09	0.082	0.098	<0.000
Uniform B (Anx all)		33	215.5	27	<0.000	0.996	0.067	0.058	0.075	<0.000
Uniform C (Mob \$3+4)		35	145.2	25	<0.000	0.998	0.055	0.047	0.064	0.152
Uniform D (Pain\$4)		36	120.2	24	<0.000	0.998	0.05	0.042	0.06	0.429
Uniform E (SC\$1)		37	109.8	23	<0.000	0.998	0.049	0.04	0.058	0.555
Uniform F (UA\$1)		38	106.4	22	<0.000	0.998	0.049	0.04	0.059	0.527
Uniform G (UA\$2)		39	99.0	21	<0.000	0.998	0.049	0.039	0.058	0.578
Uniform H (SC\$2)		40	93.8	20	<0.000	0.998	0.048	0.039	0.058	0.585
Uniform I (SC\$3)		41	83.5	19	<0.000	0.999	0.046	0.036	0.057	0.702
Uniform J (SC\$4)		42	80.6	18	<0.000	0.999	0.047	0.037	0.058	0.663
Residual Free A		45	103.4	15	<0.000	0.998	0.061	0.05	0.073	0.045
Residual Fixed B		42	80.6	18	<0.000	0.999	0.047	0.037	0.058	0.663
Factor var free		41	64.9	19	<0.000	0.999	0.039	0.028	0.05	0.953
										0.2479

Model							DIFFTEST p-value		
							RMSEA Higher CI	RMSEA Lower CI	RMSEA p-value
Free Parameter	Chi ²	DF	Chi ² p-value	CFI		RMSEA			
Configural	106	2544.6	38 <0.000	0.984	0.145	0.014	0.015	<0.000	
Non-Uniform A	100	1913.5	44 <0.000	0.988	0.116	0.112	0.121	<0.000	<0.000
Non-Uniform B (SF8)	101	1921.0	43 <0.000	0.988	0.118	0.113	0.122	<0.000	<0.000
Non-Uniform C (SF12)	102	1947.6	42 <0.000	0.988	0.12	0.116	0.125	<0.000	0.0025
Non-Uniform D (SF10)	103	2151.3	41 <0.000	0.987	0.128	0.123	0.132	<0.000	0.0262
Non-Uniform E (SF1)	104	2290.1	40 <0.000	0.986	0.134	0.129	0.138	<0.000	0.298
Uniform A	62	3754.9	82 <0.000	0.977	0.119	0.116	0.123	<0.000	<0.000
Uniform B (sf911)	70	3130.3	74 <0.000	0.981	0.115	0.111	0.118	<0.000	<0.000
Uniform C (sf23)	74	2672.0	70 <0.000	0.984	0.109	0.105	0.112	<0.000	<0.000
Uniform D (sf12)	78	2539.0	66 <0.000	0.985	0.109	0.105	0.113	<0.000	<0.000
Uniform E (sf1\$1)	79	2470.6	65 <0.000	0.985	0.108	0.105	0.112	<0.000	<0.000
Uniform F (sf10\$1)	80	2437.4	64 <0.000	0.985	0.109	0.105	0.112	<0.000	<0.000
Uniform G (sf8)	84	2397.6	60 <0.000	0.985	0.111	0.107	0.115	<0.000	<0.000
Uniform H (sf10\$2)	85	2350.0	59 <0.000	0.986	0.111	0.107	0.115	<0.000	<0.000
Uniform I (sf45\$2)	86	2319.7	58 <0.000	0.986	0.111	0.107	0.115	<0.000	<0.000
Uniform J (sf45\$1)	87	2306.9	57 <0.000	0.986	0.112	0.108	0.116	<0.000	<0.000
Uniform K (sf45\$4)	88	2274.2	56 <0.000	0.986	0.112	0.108	0.116	<0.000	<0.000
Uniform L (sf45\$3)	89	2223.5	55 <0.000	0.987	0.112	0.108	0.116	<0.000	<0.000
Uniform M (sf1\$2)	90	2193.1	54 <0.000	0.987	0.112	0.108	0.116	<0.000	<0.000
Uniform N (sf10\$3)	91	2156.2	53 <0.000	0.987	0.112	0.108	0.116	<0.000	0.0008
Uniform O (sf1\$4)	92	2167.8	52 <0.000	0.987	0.114	0.11	0.118	<0.000	0.0158
Uniform P (sf45\$7)	93	2148.3	51 <0.000	0.987	0.114	0.11	0.118	<0.000	0.1741
Residual Free	97	2555.3	47 <0.000	0.984	0.13	0.126	0.134	<0.000	
Residual Fixed	93	2148.3	51 <0.000	0.987	0.114	0.11	0.118	<0.000	0.2519

Factor variance free	91	1779.7	53	<0.000	0.989	0.102	0.102	0.098	<0.000	<0.000
----------------------	----	--------	----	--------	-------	-------	-------	-------	--------	--------

ASCOT

Model	Free Parameter	Chi ²	Chi ² DF	p-value	CFI	RMSEA	RMSEA Lower CI	RMSEA Higher CI	RMSEA p-value	DIFFTEST p-value
Configural	64	3733.8	40	<0.000	0.976	0.073	0.071	0.075	<0.000	
Non-uniform A	57	3127.6	47	<0.000	0.980	0.062	0.06	0.064	<0.000	<0.000
Non-uniform B (A1)	58	2889.8	46	<0.000	0.981	0.06	0.058	0.062	<0.000	<0.000
Non-uniform C (A4)	59	2829.3	45	<0.000	0.982	0.06	0.058	0.062	<0.000	<0.000
Non-uniform D (A7)	60	2905.6	44	<0.000	0.981	0.062	0.06	0.063	<0.000	0.011
Non-uniform E (A2)	61	2993.1	43	<0.000	0.981	0.063	0.061	0.065	<0.000	0.168
Uniform A	38	4950.0	66	<0.000	0.968	0.066	0.064	0.067	<0.000	<0.000
Uniform B (A6 all)	41	3963.2	63	<0.000	0.975	0.06	0.058	0.062	<0.000	<0.000
Uniform C (A8\$3)	42	3737.6	62	<0.000	0.976	0.059	0.057	0.06	<0.000	<0.000
Uniform D (A2\$3)	43	3478.0	61	<0.000	0.978	0.057	0.056	0.059	<0.000	<0.000
Uniform E (A1 all)	46	3255.7	58	<0.000	0.979	0.057	0.055	0.058	<0.000	<0.000
Uniform F (A5\$3)	47	3106.7	57	<0.000	0.980	0.056	0.054	0.058	<0.000	<0.000
Uniform G (A7\$3)	48	3036.6	56	<0.000	0.981	0.056	0.054	0.057	<0.000	<0.000
Uniform H (A3\$3)	49	2967.6	55	<0.000	0.981	0.056	0.054	0.057	<0.000	<0.000
Uniform I (A2 all)	51	2935.6	53	<0.000	0.981	0.056	0.055	0.058	<0.000	<0.000
Uniform J (A5\$1)	52	2924.4	52	<0.000	0.981	0.057	0.055	0.059	<0.000	<0.000
Uniform K (A3\$1)	53	2935.1	51	<0.000	0.981	0.057	0.056	0.059	<0.000	<0.000
Uniform L (A8\$1)	54	2949.5	50	<0.000	0.981	0.058	0.056	0.06	<0.000	0.001
Uniform M (A7 all)	56	2957.0	48	<0.000	0.981	0.059	0.058	0.061	<0.000	0.665
Residual Free	59	294.9	45	<0.000	0.979	0.065	0.063	0.067	<0.000	
Residual Fixed	56	2957.0	48	<0.000	0.981	0.059	0.058	0.061	<0.000	0.259
Factor var free	55	2706.2	49	<0.000	0.983	0.056	0.054	0.058	<0.000	<0.000

WEMWBS

Model	Free Parameter	Chi ²	Chi ² DF	p-value	CFI	RMSEA Lower CI	RMSEA Higher CI	RMSEA p-value	DIFFTEST p-value
Configural	142	4231.5	152	<0.000	0.958	0.123	0.12	0.126	<0.000
Non-uniform A	130	2740.4	164	<0.000	0.974	0.094	0.091	0.097	<0.000
Non-uniform B (wem2)	131	2773.1	163	<0.000	0.973	0.095	0.092	0.098	<0.000
Non-uniform C (wem4)	132	2861.1	162	<0.000	0.973	0.097	0.094	0.1	<0.000
Non-uniform D (wem11)	133	2915.7	161	<0.000	0.972	0.098	0.095	0.101	<0.000
Uniform A	79	3186.8	215	<0.000	0.97	0.088	0.085	0.091	<0.000
Uniform B (wem5)	83	3112.8	211	<0.000	0.97	0.088	0.085	0.091	<0.000
Uniform C (wem1)	87	3048.5	207	<0.000	0.971	0.088	0.085	0.091	<0.000
Uniform D (wem3 \$2+\$4)	89	3012.5	205	<0.000	0.971	0.088	0.085	0.09	<0.000
Uniform E (wem4 \$2+\$4)	91	2979.4	203	<0.000	0.972	0.088	0.085	0.09	<0.000
Uniform F (wem7\$4)	92	2958.8	202	<0.000	0.972	0.088	0.085	0.09	<0.000
Uniform G (wem3\$3)	93	2941.7	201	<0.000	0.972	0.087	0.085	0.09	<0.000
Uniform H (wem6\$4)	94	2928.6	200	<0.000	0.972	0.088	0.085	0.09	<0.000
Uniform I (wem9\$4)	95	2918.0	199	<0.000	0.972	0.088	0.085	0.09	<0.000
Uniform J (wem8\$3)	96	2906.8	198	<0.000	0.972	0.088	0.085	0.09	<0.000
Uniform K (wem13\$3)	97	2897.2	197	<0.000	0.972	0.088	0.085	0.091	<0.000
Uniform L (wem2\$3)	98	2885.4	196	<0.000	0.973	0.088	0.085	0.091	<0.000
Uniform M (wem13\$2)	99	2881.1	195	<0.000	0.973	0.088	0.085	0.091	<0.000
Uniform N (wem10\$3)	100	2873.1	194	<0.000	0.973	0.088	0.085	0.091	<0.000
Uniform O (wem13\$1)	101	2879.8	193	<0.000	0.973	0.088	0.086	0.091	<0.000
Uniform P (wem11\$4)	102	2873.6	192	<0.000	0.973	0.089	0.086	0.091	<0.000
Uniform Q (wem14\$4)	103	2871.1	191	<0.000	0.973	0.089	0.086	0.092	<0.000
									0.0004

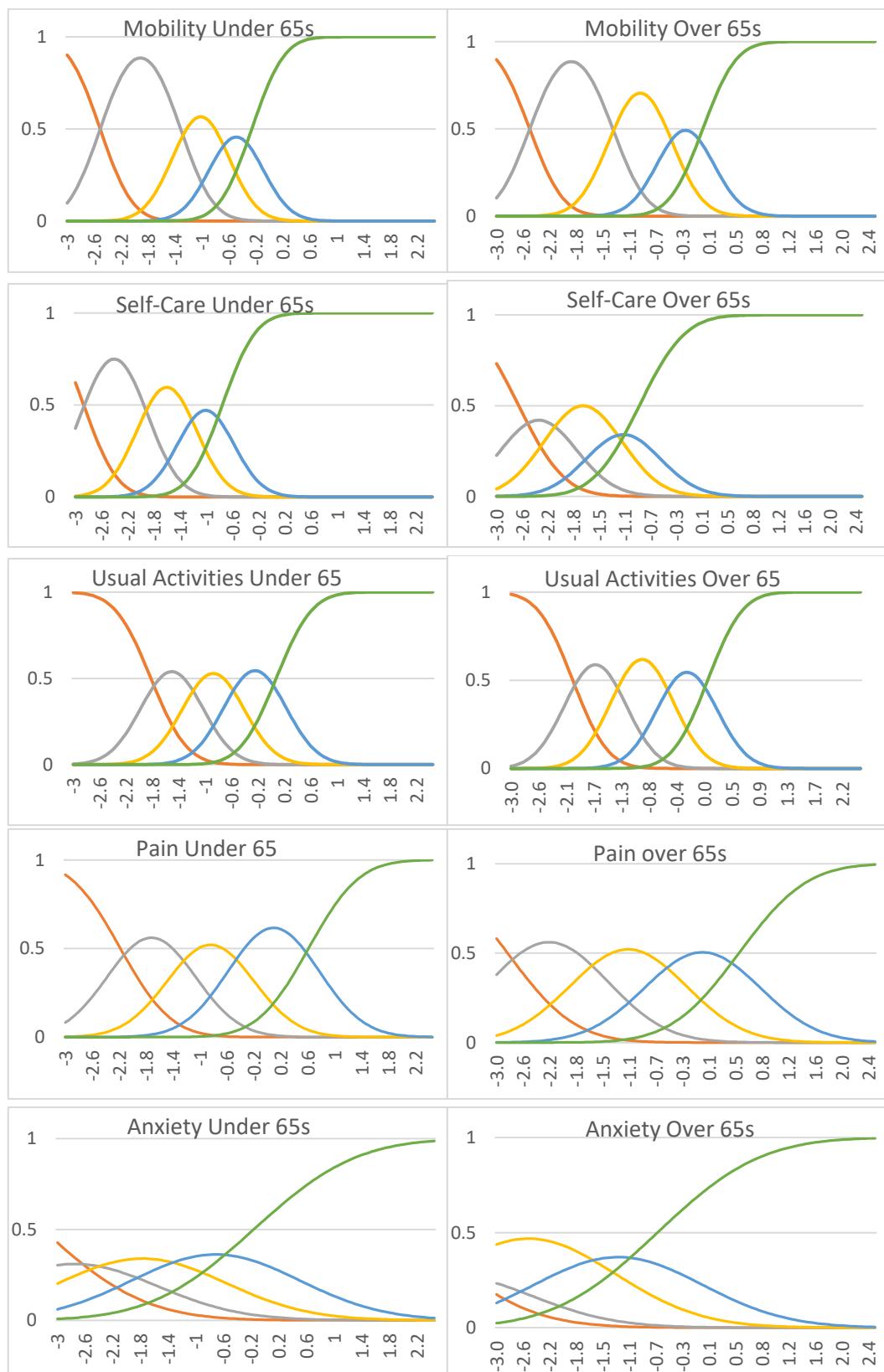
Uniform R (wem8\$4)	104	2866.3	190	<0.000	0.973	0.089	0.086	0.092	<0.000	0.0023
Uniform S (wem10 \$1+\$2+\$3)	107	2874.1	187	<0.000	0.973	0.09	0.087	0.093	<0.000	0.009
Uniform T (wem11\$2)	108	2866.7	186	<0.000	0.973	0.09	0.087	0.093	<0.000	0.0534
Residual Free	119	3871.1	175	<0.000	0.962	0.109	0.106	0.112	<0.000	
Residual Fixed	108	2866.7	186	<0.000	0.973	0.09	0.087	0.093	<0.000	0.0358
Factor var free	107	2092.1	187	<0.000	0.981	0.076	0.073	0.079	<0.000	0.005

ONS-4

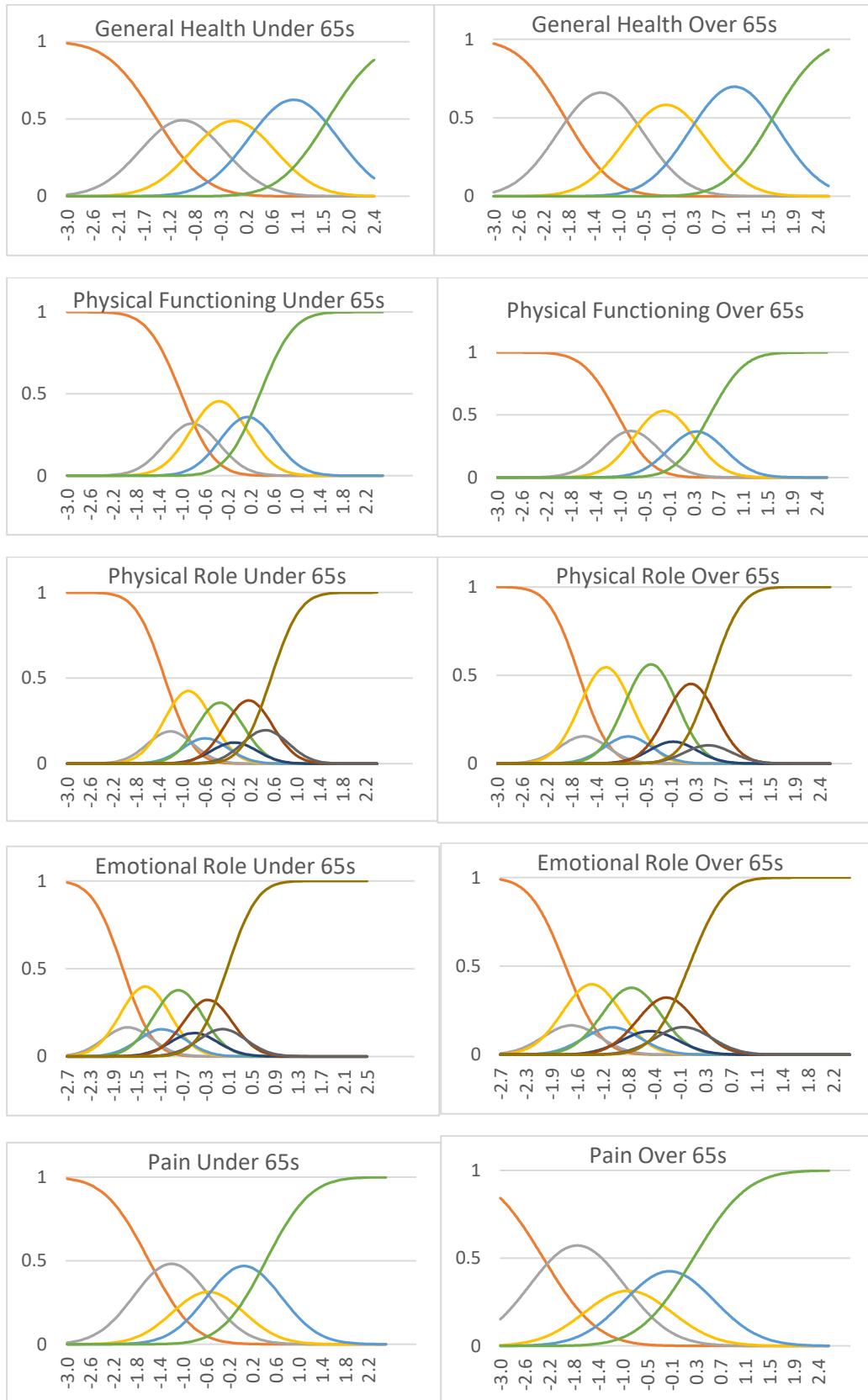
Model									DIFFTEST p-value	
	Free Parameter	Chi ²	Chi ² DF	p-value	CFI	RMSEA	RMSEA Lower CI	RMSEA Higher CI	RMSEA p-value	
Configural	80	132.1	4	<0.000	0.998	0.143	0.122	0.164	<0.000	
Non-uniform A	77	88.7	7	<0.000	0.999	0.086	0.071	0.103	<0.000	0.001
Non-uniform B (anx)	78	116.0	6	<0.000	0.998	0.108	0.091	0.126	<0.000	0.186
Uniform A	43	234.4	41	<0.000	0.997	0.055	0.048	0.062	0.117	<0.000
Uniform B (happy all)	52	177.3	32	<0.000	0.998	0.054	0.046	0.062	0.2	<0.000
Uniform C (worth\$9)	53	164.6	31	<0.000	0.998	0.052	0.045	0.06	0.294	<0.000
Uniform D (Lsat\$9)	54	152.5	30	<0.000	0.998	0.051	0.043	0.059	0.403	<0.000
Uniform E (Lsat\$5)	55	138.9	29	<0.000	0.998	0.049	0.041	0.057	0.552	<0.000
Uniform F (anx\$9)	56	133.4	28	<0.000	0.999	0.049	0.041	0.057	0.564	<0.000
Uniform G (Lsat\$8)	57	126.1	27	<0.000	0.999	0.048	0.04	0.057	0.608	<0.000
Uniform H (anx all)	65	122.5	19	<0.000	0.999	0.059	0.049	0.069	0.065	<0.000
Uniform I (worth\$8)	66	107.2	18	<0.000	0.999	0.056	0.046	0.067	0.149	0.002
Uniform J (worth\$7)	67	99.6	17	<0.000	0.999	0.056	0.045	0.066	0.178	0.019
Uniform K (Lsat\$7)	68	91.5	16	<0.000	0.999	0.055	0.044	0.066	0.219	0.157
Residual Free	71	99.8	13	<0.000	0.999	0.065	0.054	0.077	0.016	
Residual Fixed	68	91.5	16	<0.000	0.999	0.055	0.044	0.055	0.219	0.159
Factor var free	67	171.9	17	<0.000	0.998	0.076	0.066	0.087	<0.000	<0.000

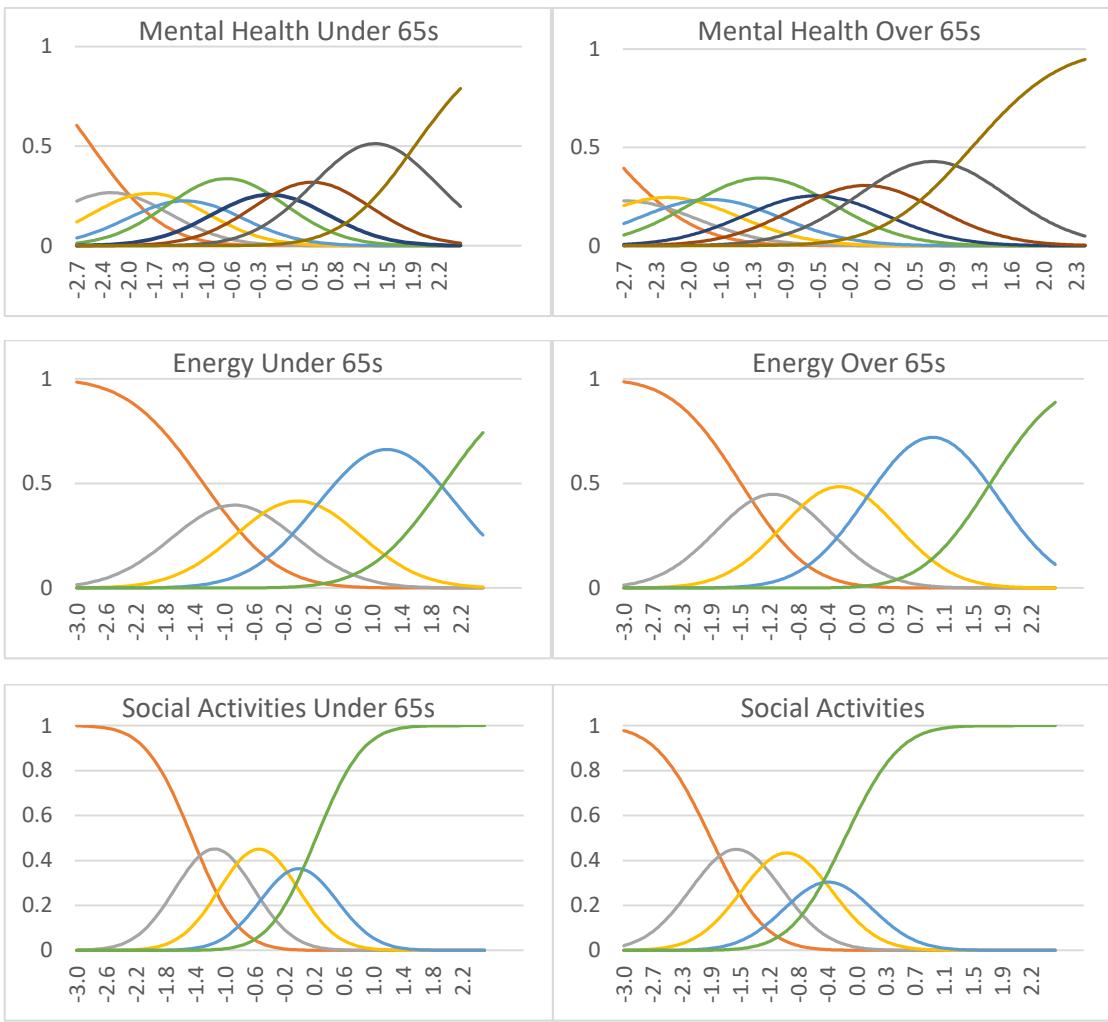
Appendix 16 – ICCs for all items

EQ-5D-5L ICCs



SF-12v2 TLA ICCs



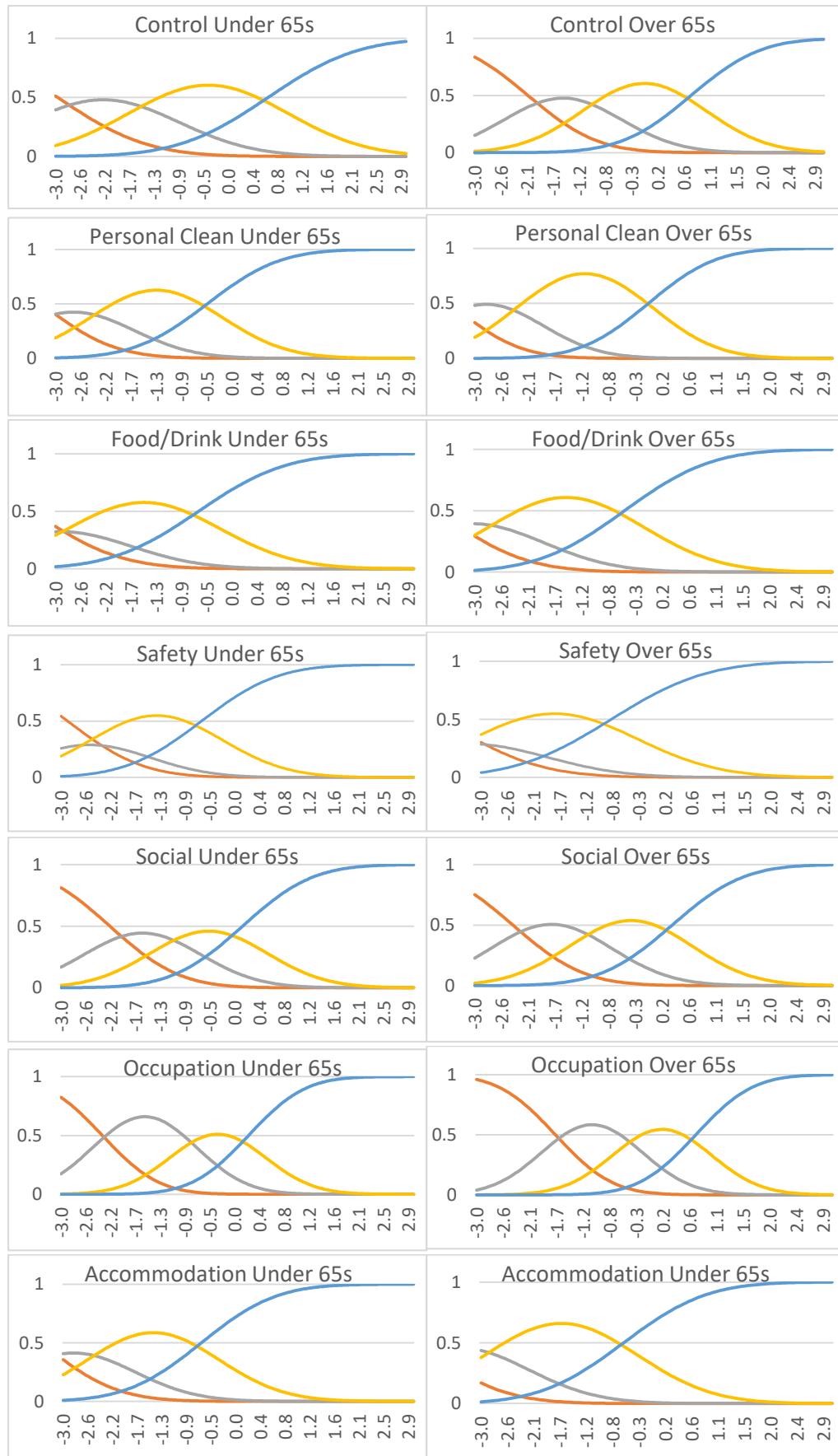


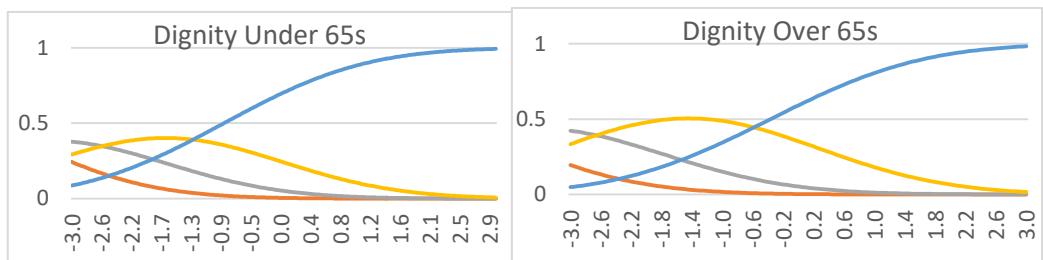
WEMWBS ICCs



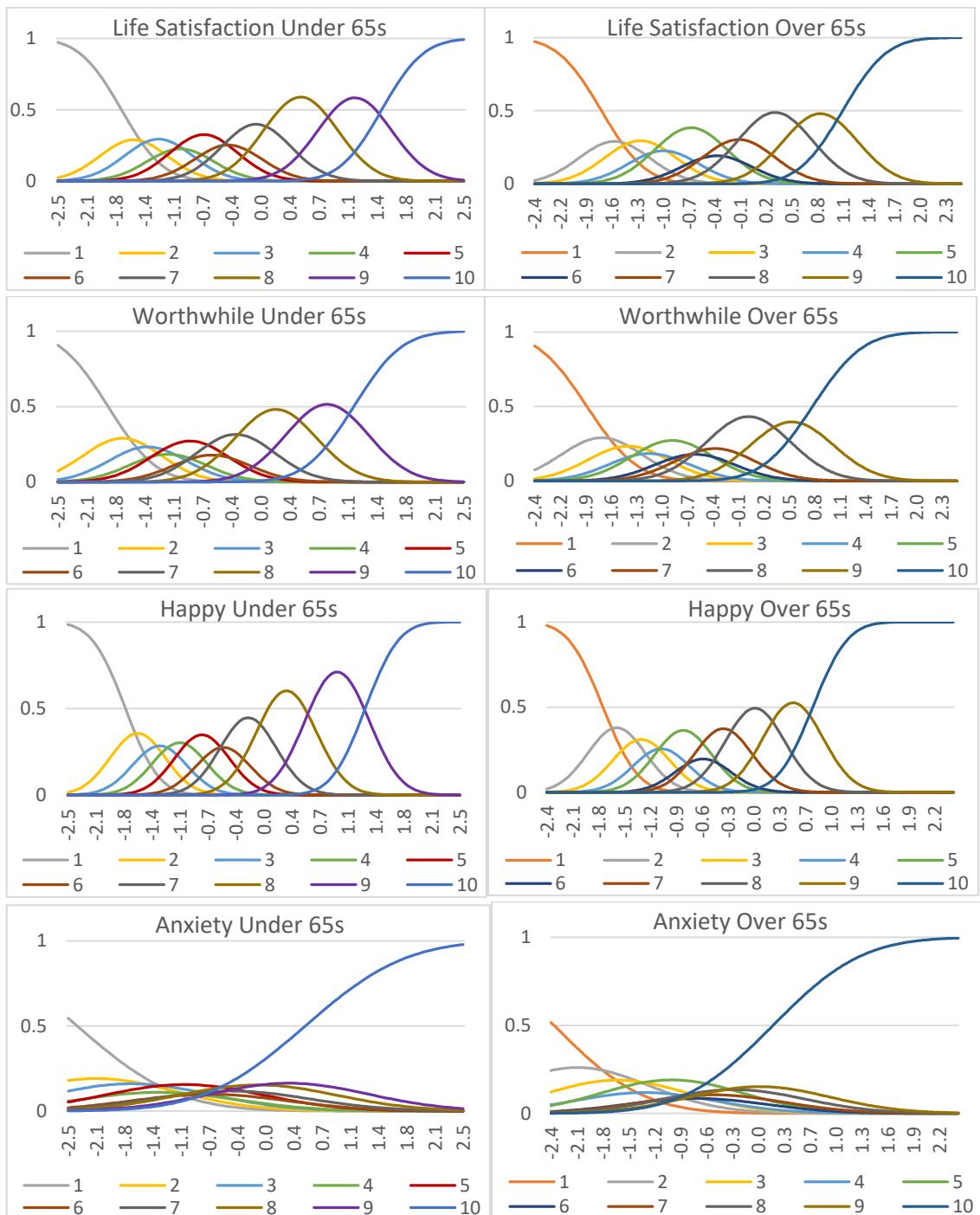


ASCOT ICCs





ONS-4 ICCs



Appendix 17 – Expected item and measure scores by age group

EQ-5D-5L Expected item decrements and utility score for under and over 65s

HRQoL	Mobility			Self-care			Usual Activities			Pain			Anxiety			Total Expected Score		
	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-3	0.27	0.27	0.00	0.19	0.19	0.00	0.18	0.18	0.00	0.33	0.30	-0.03	0.24	0.17	-0.06	-0.20	-0.12	0.09
-2.5	0.24	0.24	0.00	0.16	0.16	0.00	0.18	0.18	0.00	0.31	0.27	-0.04	0.20	0.14	-0.06	-0.10	0.01	0.11
-2	0.21	0.21	0.00	0.13	0.12	-0.01	0.17	0.17	-0.01	0.28	0.22	-0.05	0.17	0.11	-0.06	0.05	0.17	0.12
-1.5	0.17	0.17	0.00	0.08	0.08	0.00	0.14	0.13	-0.01	0.22	0.17	-0.05	0.13	0.08	-0.05	0.26	0.38	0.11
-1	0.10	0.10	0.00	0.04	0.04	0.00	0.09	0.08	-0.01	0.14	0.12	-0.03	0.09	0.06	-0.04	0.52	0.60	0.07
-0.5	0.05	0.06	0.01	0.01	0.01	0.00	0.06	0.05	0.00	0.09	0.08	-0.01	0.07	0.04	-0.03	0.72	0.75	0.03
0	0.01	0.03	0.01	0.00	0.00	0.00	0.03	0.03	0.00	0.06	0.06	-0.01	0.04	0.02	-0.02	0.85	0.86	0.01
0.5	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.04	0.03	0.00	0.03	0.01	-0.01	0.93	0.94	0.01
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.01	0.01	-0.01	0.97	0.98	0.01
1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.99	0.99	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
2.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00

SF-12v2 TLA Expected item and total scores for under and over 65s

	SF1 - Gen Health			SF2/3 - Physical functioning			SF4/5 - Physical role			SF6/7 - Emotional role			SF8 - Pain		
Health	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-3	1.01	1.03	0.02	2.00	2.00	0.00	2.00	2.00	0.00	2.00	2.00	0.00	1.01	1.18	0.17
-2.5	1.06	1.14	0.08	2.00	2.00	0.00	2.00	2.02	0.02	2.04	2.04	0.00	1.07	1.43	0.36
-2	1.22	1.42	0.20	2.02	2.02	0.00	2.05	2.31	0.26	2.43	2.43	0.00	1.27	1.81	0.54
-1.5	1.54	1.80	0.26	2.18	2.16	-0.02	2.49	3.27	0.77	3.58	3.57	-0.01	1.71	2.29	0.58
-1	2.00	2.22	0.23	2.75	2.68	-0.07	3.72	4.58	0.86	5.23	5.23	0.00	2.35	2.87	0.52
-0.5	2.52	2.67	0.15	3.69	3.50	-0.19	5.42	5.90	0.48	7.04	7.06	0.01	3.11	3.50	0.39
0	3.06	3.15	0.09	4.68	4.38	-0.30	7.29	7.35	0.06	8.77	8.76	-0.01	3.85	4.10	0.25
0.5	3.56	3.60	0.05	5.51	5.22	-0.28	8.94	8.84	-0.10	9.76	9.76	0.00	4.45	4.56	0.11
1	4.00	4.01	0.01	5.91	5.80	-0.11	9.81	9.78	-0.03	9.98	9.98	0.00	4.81	4.84	0.02
1.5	4.40	4.40	0.00	5.99	5.98	-0.01	9.99	9.99	0.00	10.00	10.00	0.00	4.96	4.96	0.00
2	4.72	4.73	0.01	6.00	6.00	0.00	10.00	10.00	0.00	10.00	10.00	0.00	5.00	4.99	0.00
2.5	4.91	4.92	0.02	6.00	6.00	0.00	10.00	10.00	0.00	10.00	10.00	0.00	5.00	5.00	0.00

	SF9/11 - Mental health			SF10 - Energy			SF12 - Social activities			Total expected physical score			Total expected mental score		
Health	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-3	2.16	2.83	0.67	1.02	1.02	0.00	1.00	1.03	0.03	8.04	8.25	0.21	4.16	4.83	0.67
-2.5	2.86	3.53	0.67	1.07	1.08	0.01	1.02	1.16	0.14	8.22	8.84	0.62	4.90	5.57	0.67
-2	3.58	4.42	0.84	1.21	1.28	0.06	1.14	1.53	0.39	8.92	10.37	1.46	6.01	6.85	0.84
-1.5	4.45	5.33	0.89	1.50	1.64	0.14	1.55	2.15	0.60	10.96	13.31	2.34	8.03	8.90	0.88
-1	5.36	6.24	0.88	1.92	2.15	0.23	2.26	2.93	0.68	14.99	17.43	2.44	10.59	11.46	0.88
-0.5	6.26	7.10	0.84	2.43	2.69	0.26	3.10	3.76	0.66	20.27	22.02	1.75	13.30	14.16	0.85
0	7.10	7.89	0.78	2.96	3.22	0.26	3.96	4.47	0.51	25.80	26.68	0.88	15.87	16.65	0.78
0.5	7.88	8.60	0.72	3.43	3.66	0.23	4.63	4.86	0.23	30.51	30.75	0.23	17.63	18.36	0.72
1	8.54	9.17	0.63	3.83	4.02	0.19	4.93	4.98	0.05	33.30	33.43	0.14	18.52	19.15	0.63
1.5	9.08	9.59	0.52	4.18	4.36	0.18	4.99	5.00	0.00	34.51	34.68	0.17	19.07	19.59	0.52
2	9.49	9.84	0.35	4.48	4.66	0.18	5.00	5.00	0.00	35.20	35.38	0.19	19.49	19.84	0.35
2.5	9.80	9.97	0.17	4.74	4.88	0.14	5.00	5.00	0.00	35.64	35.80	0.16	19.80	19.97	0.17

ASCOT Expected item and total scores for under and over 65s

SCRQoL	1. Control			2. Personal clean/comfort			3. Food/drink			4. safety			5. social participation		
	65+	18-64	diff	65+	18-64	diff	65+	18-64	diff	65+	18-64	diff	65+	18-64	diff
-3	0.09	0.30	-0.21	0.34	0.34	0.00	0.41	0.41	0.01	0.32	0.23	0.09	0.31	0.30	0.01
-2.5	0.19	0.42	-0.23	0.45	0.44	0.01	0.51	0.51	0.00	0.39	0.30	0.09	0.37	0.35	0.02
-2	0.33	0.55	-0.21	0.58	0.56	0.01	0.61	0.61	0.00	0.46	0.39	0.08	0.45	0.43	0.02
-1.5	0.49	0.66	-0.17	0.69	0.68	0.01	0.69	0.69	0.00	0.54	0.48	0.06	0.54	0.52	0.01
-1	0.65	0.76	-0.11	0.76	0.76	0.00	0.76	0.76	0.00	0.62	0.58	0.04	0.62	0.62	0.00
-0.5	0.77	0.84	-0.07	0.81	0.83	-0.01	0.80	0.81	0.00	0.69	0.68	0.02	0.70	0.70	-0.01
0	0.86	0.89	-0.03	0.85	0.87	-0.01	0.84	0.84	0.00	0.76	0.76	0.00	0.76	0.77	-0.01
0.5	0.92	0.93	-0.01	0.88	0.89	-0.01	0.86	0.86	0.00	0.81	0.82	-0.01	0.81	0.82	-0.01
1	0.96	0.96	0.00	0.90	0.90	-0.01	0.87	0.87	0.00	0.84	0.86	-0.01	0.84	0.85	-0.01
1.5	0.98	0.98	0.00	0.91	0.91	0.00	0.87	0.88	0.00	0.86	0.87	-0.01	0.86	0.86	0.00
2	0.99	0.99	0.00	0.91	0.91	0.00	0.88	0.88	0.00	0.87	0.88	0.00	0.87	0.87	0.00
2.5	1.00	0.99	0.00	0.91	0.91	0.00	0.88	0.88	0.00	0.88	0.88	0.00	0.87	0.87	0.00

	6. Occupation			7. Accommodation clean/comf			8. Dignity			Total Expected Score		
SCRQoL	65+	18-64	diff	65+	18-64	diff	65+	18-64	diff	65+	18-64	diff
-3	0.31	0.30	0.01	0.52	0.44	0.08	0.43	0.44	-0.01	0.09	0.09	0.00
-2.5	0.37	0.35	0.02	0.60	0.53	0.07	0.48	0.49	-0.01	0.22	0.23	-0.01
-2	0.45	0.43	0.02	0.68	0.62	0.05	0.54	0.56	-0.02	0.37	0.38	-0.01
-1.5	0.54	0.52	0.01	0.74	0.71	0.03	0.60	0.62	-0.02	0.51	0.53	-0.01
-1	0.62	0.62	0.00	0.78	0.77	0.01	0.65	0.68	-0.02	0.64	0.66	-0.02
-0.5	0.70	0.70	-0.01	0.81	0.81	0.00	0.70	0.73	-0.03	0.75	0.77	-0.02
0	0.76	0.77	-0.01	0.83	0.84	0.00	0.74	0.77	-0.02	0.83	0.85	-0.02
0.5	0.81	0.82	-0.01	0.85	0.85	0.00	0.78	0.80	-0.02	0.90	0.91	-0.02
1	0.84	0.85	-0.01	0.86	0.86	0.00	0.80	0.82	-0.02	0.94	0.95	-0.01
1.5	0.86	0.86	0.00	0.86	0.86	0.00	0.82	0.83	-0.01	0.96	0.97	-0.01
2	0.87	0.87	0.00	0.86	0.86	0.00	0.83	0.84	-0.01	0.97	0.97	0.00
2.5	0.87	0.87	0.00	0.86	0.86	0.00	0.84	0.84	0.00	0.98	0.98	0.00

WEMWBS Expected item and total scores for under and over 65s

	1. Optimistic			2. Useful			3. Relaxed			4. Interest other people			5. Energy to Spare			6. Deal problems well		
wellbeing	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-2.5	1.78	1.49	-0.29	1.70	1.24	-0.46	1.59	1.70	0.12	1.95	1.77	-0.17	1.56	1.35	-0.21	1.72	1.72	0.00
-2	2.10	1.77	-0.33	2.12	1.61	-0.51	1.93	2.14	0.20	2.30	2.18	-0.11	1.82	1.58	-0.23	2.21	2.21	0.00
-1.5	2.45	2.11	-0.34	2.56	2.14	-0.42	2.30	2.56	0.26	2.65	2.60	-0.05	2.09	1.87	-0.23	2.67	2.67	0.00
-1	2.79	2.46	-0.33	2.97	2.67	-0.29	2.66	2.91	0.25	2.99	2.98	-0.01	2.38	2.18	-0.21	3.04	3.04	0.00
-0.5	3.12	2.82	-0.30	3.33	3.12	-0.21	3.00	3.22	0.22	3.31	3.34	0.03	2.68	2.50	-0.18	3.38	3.38	0.01
0	3.42	3.17	-0.26	3.65	3.54	-0.12	3.32	3.53	0.21	3.61	3.69	0.08	2.97	2.80	-0.17	3.70	3.74	0.04
0.5	3.72	3.51	-0.21	3.96	3.97	0.01	3.65	3.85	0.21	3.89	4.03	0.14	3.27	3.10	-0.16	4.00	4.09	0.09
1	4.00	3.84	-0.16	4.24	4.39	0.14	3.96	4.17	0.21	4.15	4.35	0.20	3.56	3.39	-0.17	4.29	4.44	0.15
1.5	4.26	4.15	-0.11	4.51	4.72	0.20	4.27	4.48	0.21	4.39	4.62	0.23	3.86	3.69	-0.17	4.59	4.73	0.14
2	4.49	4.43	-0.07	4.74	4.91	0.17	4.55	4.73	0.18	4.60	4.81	0.21	4.15	3.97	-0.17	4.83	4.91	0.08
2.5	4.68	4.65	-0.04	4.89	4.98	0.09	4.78	4.89	0.11	4.77	4.92	0.15	4.41	4.25	-0.16	4.95	4.98	0.03

	7. Thinking clearly			8. Feel good self			9. Close other people			10. Confident			11. Make own mind			12. Loved		
wellbeing	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-2.5	1.87	1.86	0.00	1.40	1.42	0.03	1.68	1.68	0.00	1.37	1.61	0.24	2.27	2.12	- 0.15	1.96	1.96	0.00
-2	2.37	2.36	0.00	1.89	1.89	0.00	2.12	2.12	0.00	1.86	2.06	0.20	2.69	2.62	- 0.07	2.41	2.41	0.00
-1.5	2.82	2.82	0.00	2.41	2.39	- 0.02	2.56	2.56	0.00	2.39	2.52	0.12	3.09	3.04	- 0.05	2.87	2.87	0.00
-1	3.22	3.22	0.01	2.85	2.81	- 0.04	2.96	2.96	0.00	2.85	2.89	0.04	3.45	3.41	- 0.04	3.32	3.31	0.00
-0.5	3.58	3.61	0.03	3.18	3.13	- 0.05	3.33	3.34	0.01	3.21	3.17	- 0.04	3.78	3.79	0.01	3.74	3.74	0.00
0	3.90	4.00	0.10	3.59	3.51	- 0.09	3.68	3.71	0.03	3.63	3.57	- 0.06	4.08	4.16	0.08	4.13	4.13	0.00
0.5	4.19	4.37	0.18	3.95	3.94	- 0.01	4.02	4.09	0.07	3.98	3.99	0.01	4.37	4.51	0.14	4.47	4.47	0.00
1	4.50	4.69	0.19	4.28	4.37	0.10	4.34	4.45	0.10	4.32	4.41	0.09	4.63	4.78	0.14	4.73	4.73	0.00
1.5	4.77	4.89	0.12	4.68	4.75	0.07	4.64	4.73	0.09	4.72	4.80	0.07	4.83	4.93	0.10	4.89	4.89	0.00
2	4.93	4.98	0.05	4.93	4.94	0.01	4.86	4.91	0.05	4.95	4.97	0.02	4.94	4.98	0.04	4.96	4.96	0.00
2.5	4.99	5.00	0.01	4.99	4.99	0.00	4.96	4.98	0.02	5.00	5.00	0.00	4.98	5.00	0.01	4.99	4.99	0.00

wellbeing	13. Interest new things			14. cheerful			Total Expected Score		
	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-2.5	1.68	1.50	-0.19	1.65	1.65	0.00	24.17	23.09	-1.09
-2	2.09	1.87	-0.21	2.15	2.14	0.00	30.06	28.98	-1.08
-1.5	2.51	2.30	-0.21	2.62	2.62	0.00	36.00	35.05	-0.95
-1	2.92	2.73	-0.19	3.00	3.00	0.00	41.39	40.59	-0.81
-0.5	3.31	3.14	-0.16	3.37	3.37	0.00	46.31	45.67	-0.64
0	3.67	3.54	-0.13	3.74	3.75	0.01	51.13	50.84	-0.29
0.5	4.02	3.93	-0.09	4.02	4.08	0.06	55.51	55.92	0.41
1	4.35	4.31	-0.04	4.31	4.43	0.11	59.67	60.74	1.07
1.5	4.63	4.61	-0.01	4.67	4.77	0.10	63.72	64.75	1.04
2	4.83	4.82	0.00	4.91	4.95	0.04	66.67	67.28	0.60
2.5	4.94	4.94	0.00	4.99	4.99	0.01	68.32	68.56	0.24

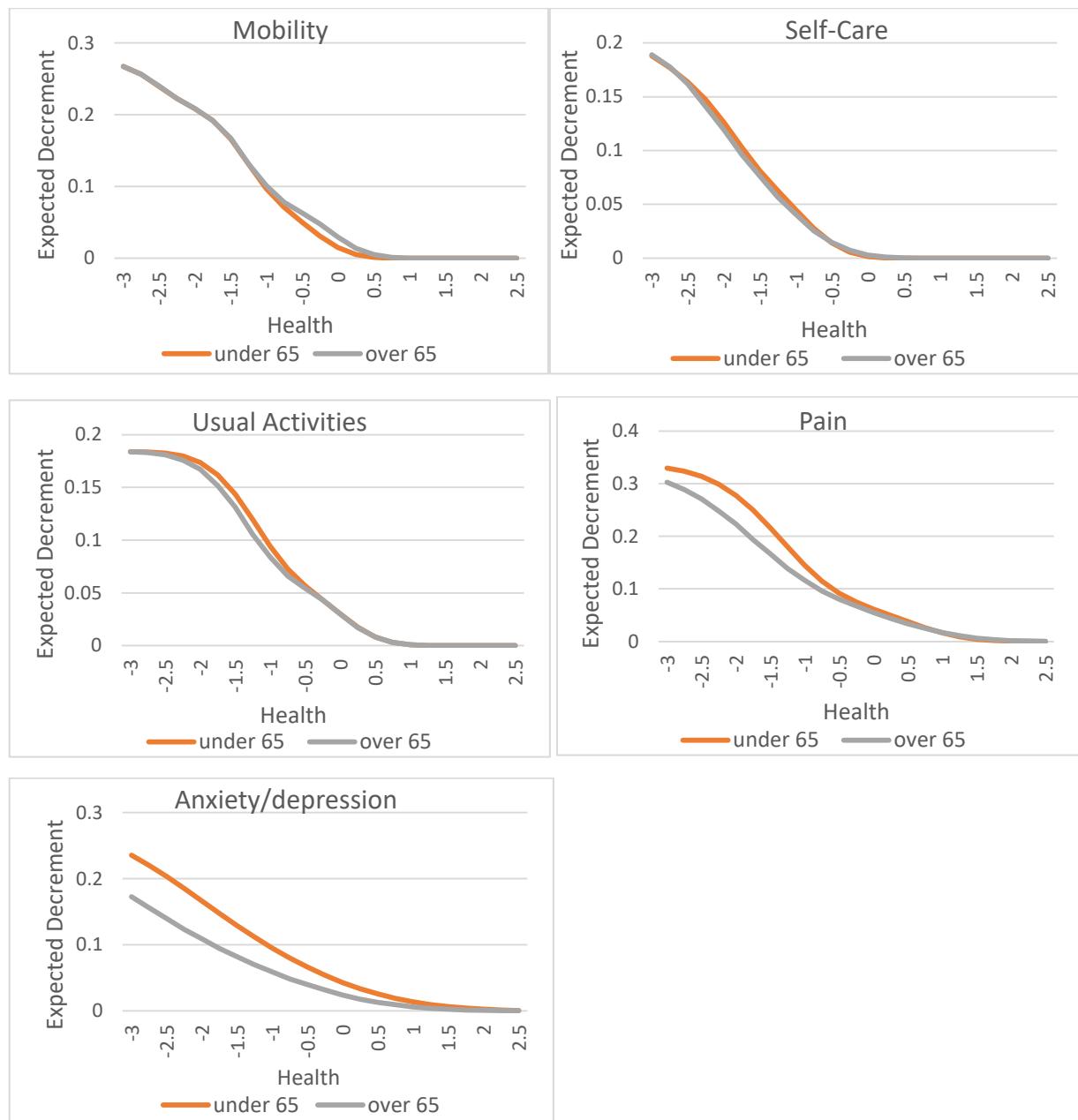
ONS-4 Expected item and total scores for under and over 65s

	Life satisfaction			Worthwhile			Happy			Anxiety			Total Expected Score		
wellbeing	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff	18-64	65+	Diff
-2.25	1.09	1.09	0.00	1.26	1.25	0.00	1.06	1.08	0.02	2.39	2.26	-0.13	5.80	5.68	-0.12
-2	1.29	1.29	0.00	1.57	1.57	0.00	1.25	1.30	0.05	2.83	2.72	-0.11	6.94	6.88	-0.06
-1.5	2.27	2.27	-0.01	2.75	2.75	0.00	2.37	2.46	0.08	3.91	3.87	-0.04	11.31	11.34	0.03
-1	3.88	3.85	-0.03	4.45	4.46	0.01	4.16	4.27	0.11	5.18	5.27	0.09	17.67	17.85	0.18
-0.5	5.58	5.54	-0.04	6.17	6.26	0.10	5.96	6.21	0.26	6.49	6.75	0.26	24.19	24.76	0.57
0	7.00	7.14	0.14	7.53	7.79	0.26	7.39	7.82	0.43	7.70	8.10	0.40	29.61	30.84	1.23
0.5	8.00	8.34	0.34	8.49	8.88	0.39	8.36	8.96	0.60	8.66	9.06	0.40	33.52	35.25	1.73
1	8.79	9.27	0.48	9.26	9.63	0.37	9.11	9.76	0.65	9.32	9.63	0.31	36.48	38.29	1.80
1.5	9.50	9.84	0.34	9.78	9.94	0.16	9.75	9.99	0.23	9.71	9.88	0.17	38.74	39.65	0.90
2	9.91	9.99	0.08	9.97	10.00	0.02	9.99	10.00	0.01	9.90	9.97	0.08	39.77	39.96	0.19
2.25	9.98	10.00	0.02	9.99	10.00	0.01	10.00	10.00	0.00	9.94	9.99	0.05	39.91	39.99	0.08

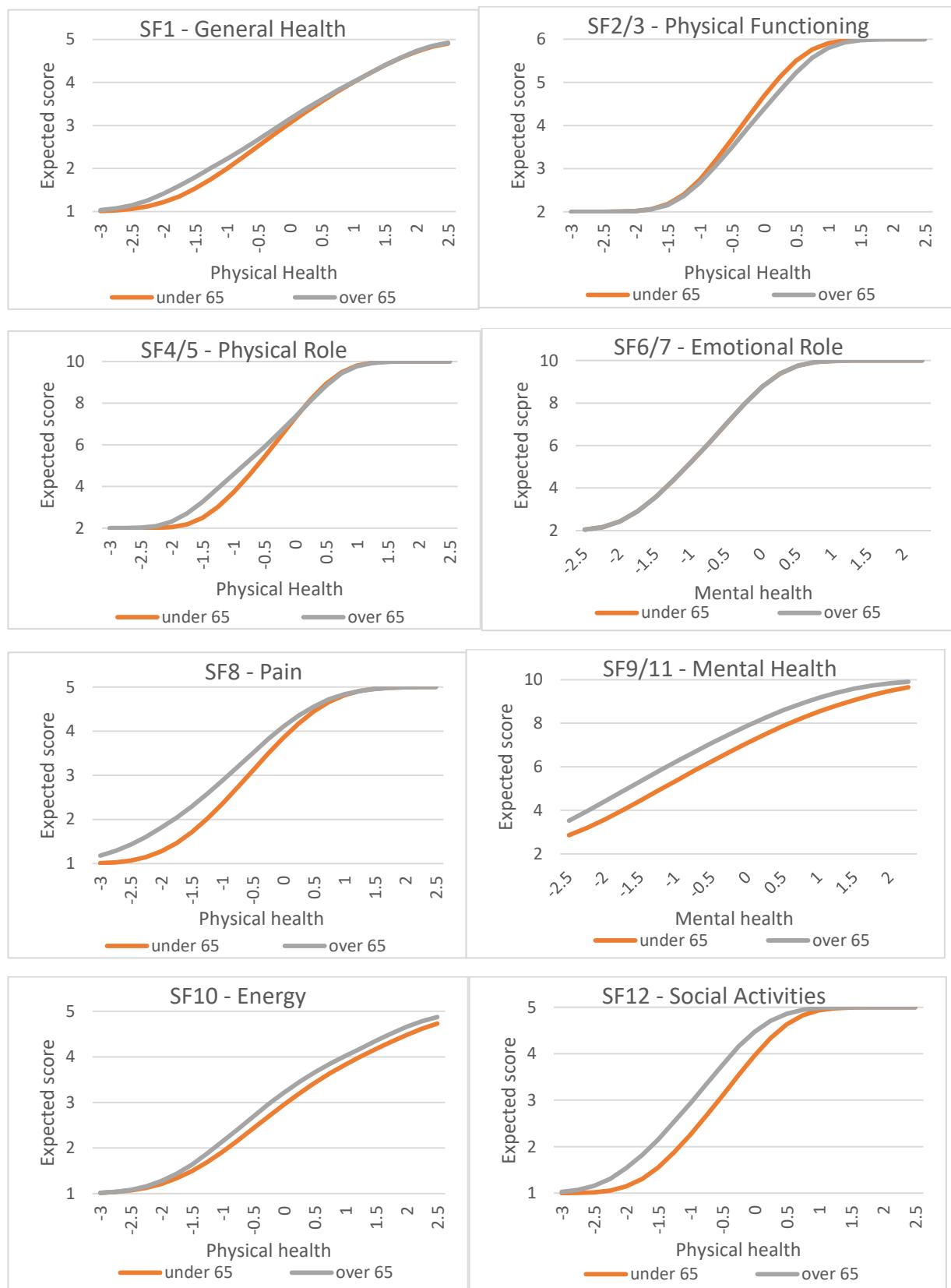
Appendix 18 - Expected item score figures for under and over 65s

EQ-5D-5L

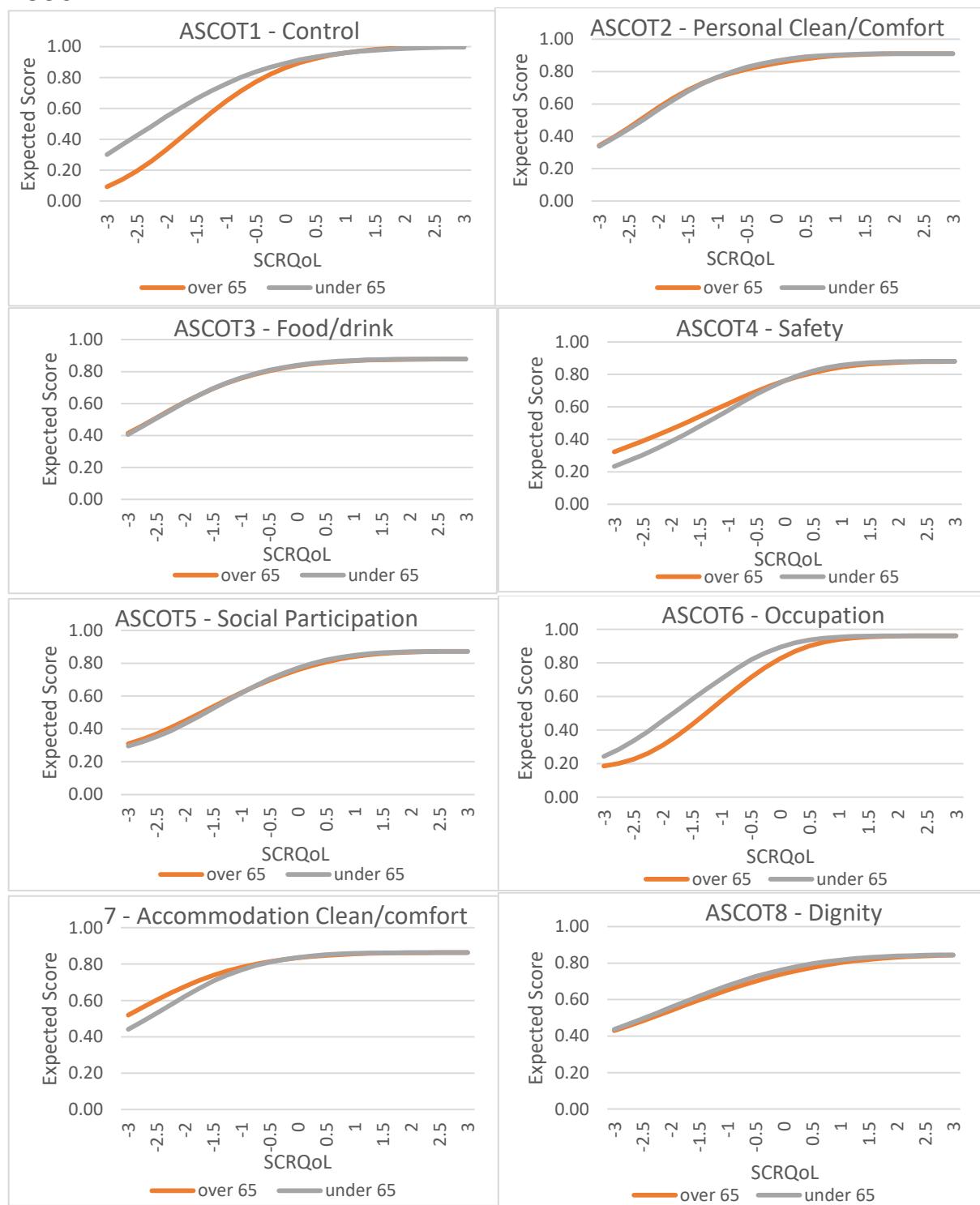
To account for the fact that the EQ-5D-5L is preference-based, each graph shows the expected decrement associated from that item at any given level of underlying health for under and over 65s.



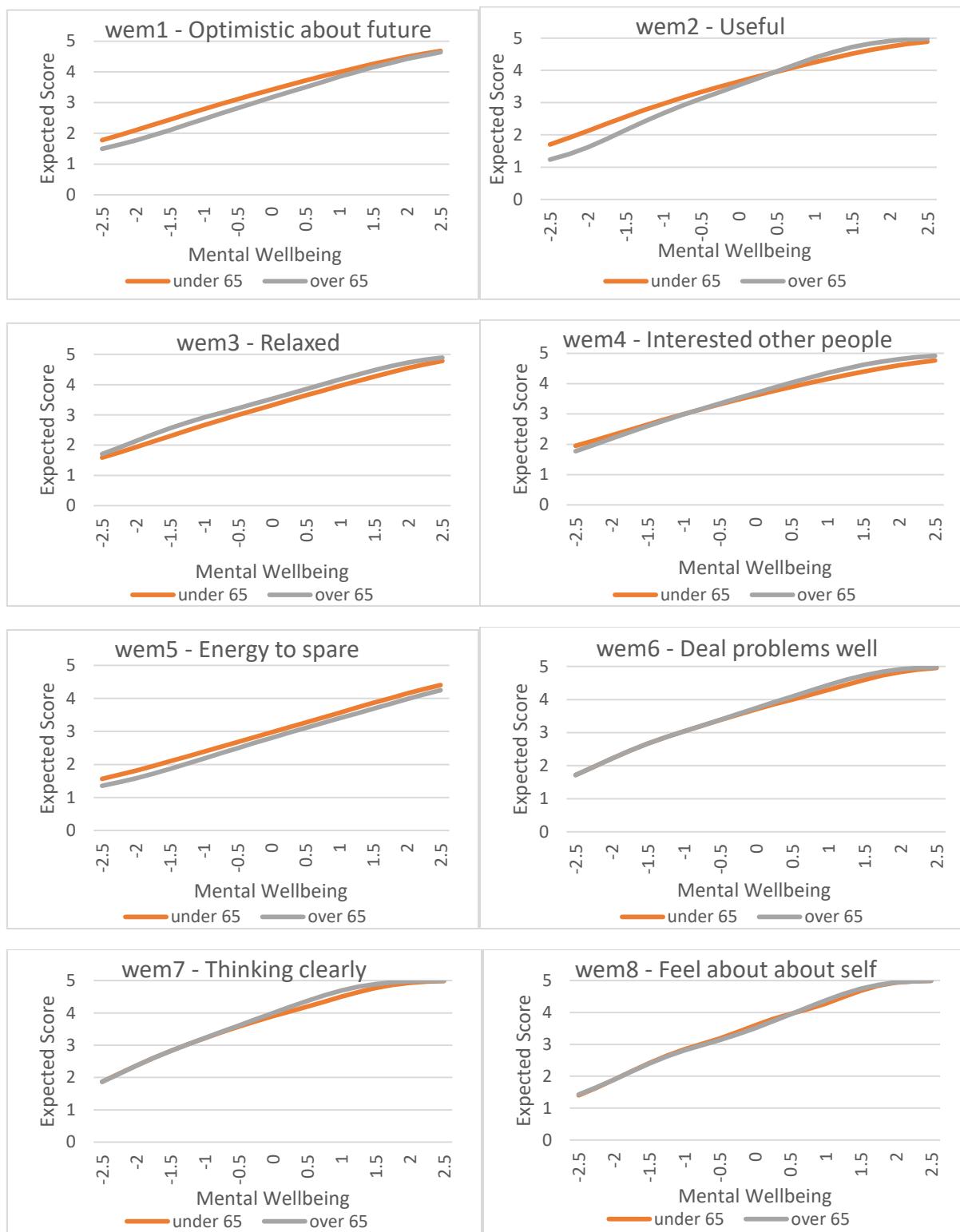
SF-12v2 TLA

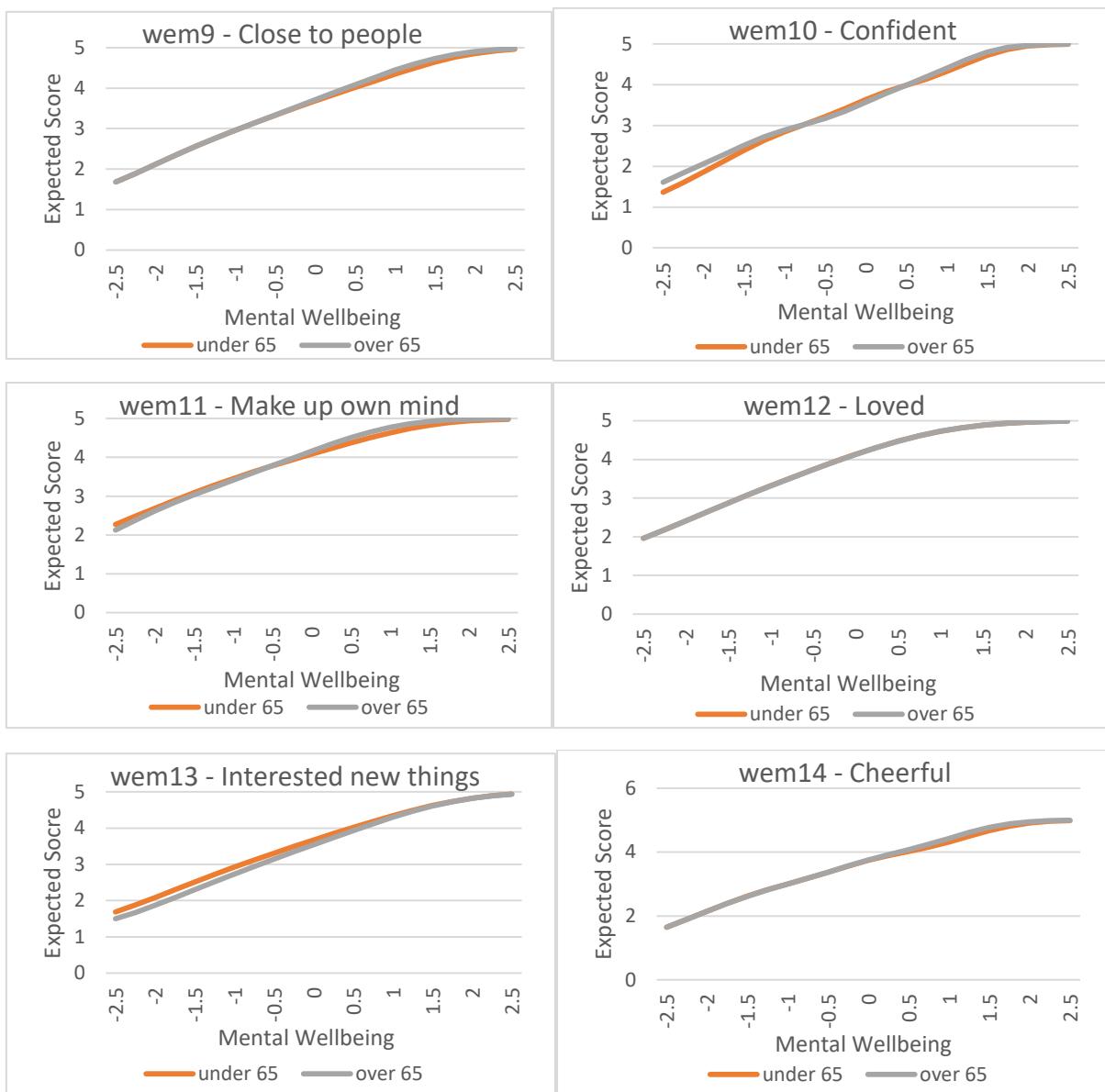


ASCOT

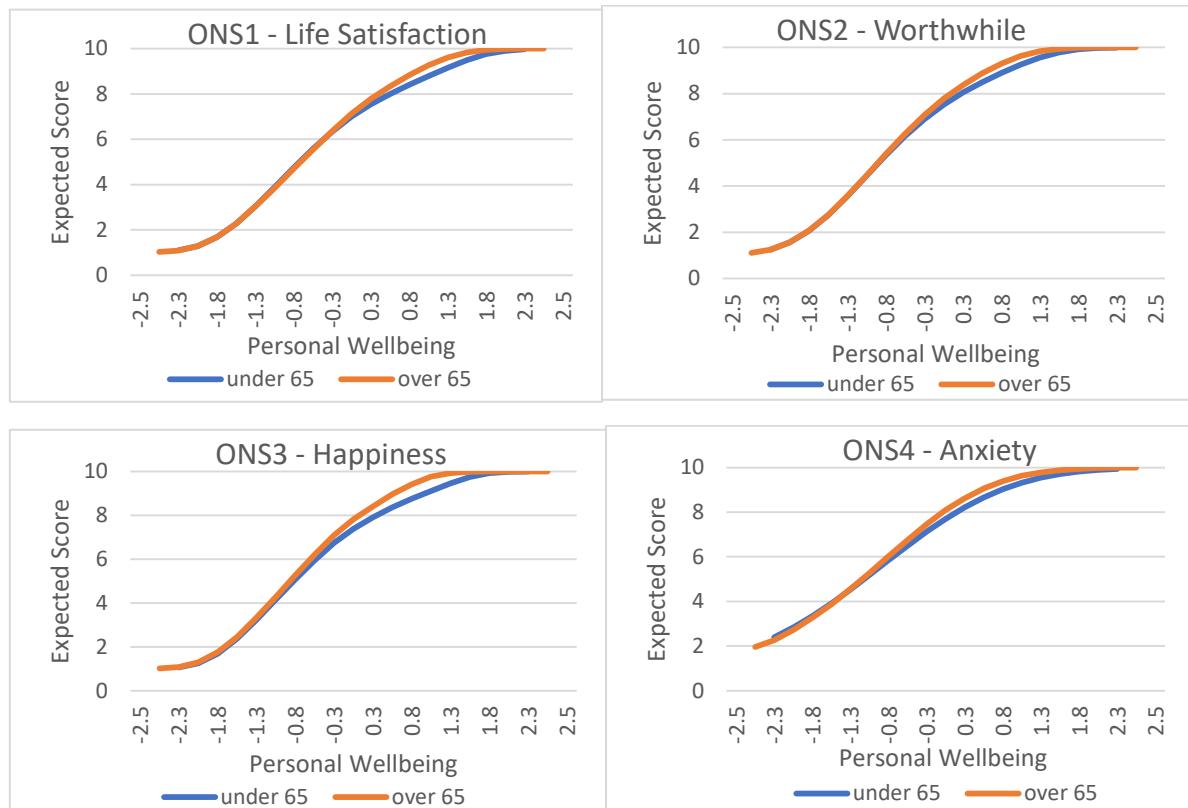


WEMWBS





ONS-4



Appendix 19 – IRT parameters from final DIF model from development and validation samples

EQ-5D-5L

Item	Para-meter	Under 65			Over 65		
		Development	Validation	Difference	Development	Validation	Difference
Mobility	a	2.64	2.655	-0.015	2.64	2.655	-0.015
	B1	-2.511	-2.525	0.014	-2.511	-2.525	0.014
	B2	-1.314	-1.347	0.033	-1.314	-1.347	0.033
	B3	-0.720	-0.789	0.069	-0.520	-0.531	0.011
	B4	-0.259	-0.277	0.017	-0.019	-0.056	0.037
Self-care	a	2.469	2.363	0.106	1.858	2.058	-0.2
	B1	-2.875	-2.500	-0.375	-2.655	-2.679	0.024
	B2	-1.943	-1.971	0.028	-2.059	-2.138	0.079
	B3	-1.266	-1.252	-0.014	-1.334	-1.362	0.027
	B4	-0.756	-0.835	0.079	-0.860	-0.788	-0.072
Usual acts	a	2.322	2.377	-0.055	2.322	2.377	-0.055
	B1	-1.823	-1.895	0.072	-2.028	-1.998	-0.030
	B2	-1.188	-1.243	0.055	-1.321	-1.377	0.056
	B3	-0.568	-0.598	0.030	-0.568	-0.598	0.030
	B4	0.076	0.057	0.019	0.076	0.057	0.019
Pain/discom	a	1.685	1.628	0.057	1.299	1.277	0.022
	B1	-2.181	-2.149	-0.032	-2.829	-2.740	-0.089
	B2	-1.262	-1.264	0.002	-1.637	-1.612	-0.026
	B3	-0.423	-0.433	0.010	-0.548	-0.552	0.004
	B4	0.612	0.552	0.060	0.501	0.527	-0.026
Anxiety/depress	a	0.857	0.857	0	0.857	0.857	0
	B1	-1.635	-1.546	-0.089	-2.692	-2.778	0.086
	B2	-1.162	-1.152	-0.010	-2.150	-2.174	0.024
	B3	-0.640	-0.607	-0.033	-1.187	-1.246	0.059
	B4	-0.081	-0.041	-0.041	-0.443	-0.399	-0.044

Item	Para-meter	Under 65			Over 65		
		Development	Validation	Difference	Development	Validation	Difference
General Health	A	1.461	1.465	0.004	1.620	1.620	0.011
	B1	-1.426	-1.416	0.009	-1.841	-1.841	-0.005
	B2	-0.520	-0.523	-0.003	-0.661	-0.661	0.034
	B3	0.377	0.390	0.013	0.340	0.340	0.011
	B4	1.591	1.593	0.002	1.615	1.615	0.009
Physical functioning	A	2.164	2.232	0.068	2.164	2.164	0.068
	B1	-1.012	-1.022	-0.009	-0.999	-0.999	0.026
	B2	-0.633	-0.642	-0.009	-0.552	-0.552	0.000
	B3	-0.073	-0.093	-0.020	0.119	0.119	0.009
	B4	0.358	0.331	-0.027	0.561	0.561	0.038
Physical role	A	2.522	2.445	-0.077	2.522	2.522	-0.077
	B1	-1.299	-1.257	0.042	-1.664	-1.664	-0.024
	B2	-1.110	-1.098	0.012	-1.510	-1.510	-0.005
	B3	-0.665	-0.660	0.005	-0.917	-0.917	0.009
	B4	-0.517	-0.516	0.001	-0.764	-0.764	0.006
	B5	-0.149	-0.141	0.008	-0.149	-0.149	0.008
	B6	-0.026	-0.024	0.002	-0.026	-0.026	0.002
	B7	0.355	0.348	-0.007	0.449	0.449	0.026
	B8	0.552	0.573	0.021	0.552	0.552	0.021
Emotional role	A	2.433	2.587	0.154	2.433	2.433	0.154
	B1	-1.734	-1.697	0.038	-1.734	-1.734	0.038
	B2	-1.562	-1.532	0.030	-1.562	-1.562	0.030
	B3	-1.134	-1.092	0.041	-1.134	-1.134	0.041
	B4	-0.973	-0.942	0.031	-0.973	-0.973	0.031
	B5	-0.568	-0.555	0.013	-0.568	-0.568	0.013
	B6	-0.430	-0.409	0.021	-0.430	-0.430	0.021
	B7	-0.088	-0.073	0.015	-0.088	-0.088	0.015
	B8	0.074	0.089	0.015	0.074	0.074	0.015
Pain	A	1.660	1.631	-0.029	1.381	1.381	0.011
	B1	-1.587	-1.584	0.004	-2.306	-2.306	-0.022
	B2	-0.810	-0.830	-0.021	-1.157	-1.157	0.032
	B3	-0.323	-0.327	-0.004	-0.571	-0.571	0.037
	B4	0.430	0.408	-0.022	0.242	0.242	0.035
Mental Health	A	1.293	1.288	-0.005	1.293	1.293	-0.005
	B1	-2.496	-2.435	0.061	-2.896	-2.896	0.042
	B2	-1.970	-1.887	0.083	-2.448	-2.448	-0.064
	B3	-1.450	-1.391	0.060	-1.964	-1.964	-0.038
	B4	-1.008	-0.973	0.035	-1.500	-1.500	0.010
	B5	-0.334	-0.286	0.048	-0.811	-0.811	0.021
	B6	0.171	0.215	0.044	-0.311	-0.311	-0.017
	B7	0.805	0.805	0.000	0.299	0.299	-0.057
	B8	1.879	1.808	-0.070	1.173	1.173	-0.037
Energy	A	1.242	1.263	0.021	1.442	1.442	0.052
	B1	-1.283	-1.261	0.022	-1.520	-1.520	-0.002
	B2	-0.448	-0.436	0.011	-0.696	-0.696	0.038
	B3	0.434	0.433	-0.001	0.205	0.205	0.000
	B4	1.977	2.004	0.027	1.703	1.703	-0.009

Social activities	B1	-1.441	-1.447	-0.006	-1.912	-1.912	-0.007
	B2	-0.842	-0.824	0.018	-1.247	-1.247	-0.011
	B3	-0.243	-0.247	-0.004	-0.609	-0.609	-0.008
	B4	0.229	0.220	-0.009	-0.175	-0.175	0.005

ASCOT

Item	Para-meter	Under 65			Over 65		
		Development	Validation	Difference	Development	Validation	Difference
Control	A	0.827	0.831	-0.004	1.065	1.043	0.022
	B1	-2.975	-3.001	0.027	-2.082	-2.121	0.039
	B2	-1.417	-1.446	0.029	-0.881	-0.887	0.006
	B3	0.636	0.625	0.011	0.720	0.716	0.004
Personal clean/ comfort	A	1.04	1.032	0.008	1.122	1.075	0.047
	B1	-3.247	-3.478	0.231	-3.402	-3.420	0.019
	B2	-2.172	-2.282	0.110	-2.223	-2.280	0.057
	B3	-0.464	-0.543	0.078	-0.083	-0.100	0.017
Food/ drink	A	0.873	0.85	0.023	0.873	0.85	0.023
	B1	-3.397	-3.561	0.164	-3.623	-3.749	0.126
	B2	-2.444	-2.496	0.052	-2.444	-2.496	0.052
	B3	-0.609	-0.672	0.062	-0.483	-0.482	-0.001
Safety	A	0.983	0.944	0.039	0.779	0.803	-0.024
	B1	-2.900	-2.995	0.094	-3.660	-3.521	-0.139
	B2	-2.149	-2.229	0.080	-2.711	-2.620	-0.091
	B3	-0.614	-0.648	0.034	-0.775	-0.762	-0.013
Social Participation	A	1.062	1.054	0.008	1.062	1.054	0.008
	B1	-2.179	-2.194	0.016	-2.359	-2.363	0.005
	B2	-1.070	-1.071	0.001	-1.070	-1.071	0.001
	B3	0.084	0.076	0.008	0.315	0.285	0.031
Occupation	A	1.319	1.349	-0.030	1.319	1.349	-0.030
	B1	-2.308	-2.309	0.001	-1.669	-1.632	-0.036
	B2	-0.862	-0.865	0.003	-0.434	-0.421	-0.013
	B3	0.187	0.165	0.021	0.700	0.695	0.005
Accommodation clean/ comfort	A	0.993	0.954	0.039	0.893	0.891	0.002
	B1	-3.382	-3.503	0.121	-4.072	-4.030	-0.041
	B2	-2.286	-2.387	0.101	-2.697	-2.704	0.007
	B3	-0.639	-0.673	0.033	-0.562	-0.575	0.012
Dignity	A	0.631	0.63	0.001	0.631	0.63	0.001
	B1	-4.117	-4.233	0.116	-4.365	-4.395	0.031
	B2	-2.528	-2.570	0.042	-2.528	-2.570	0.042
	B3	-0.856	-0.878	0.022	-0.366	-0.368	0.002

WEMWBS

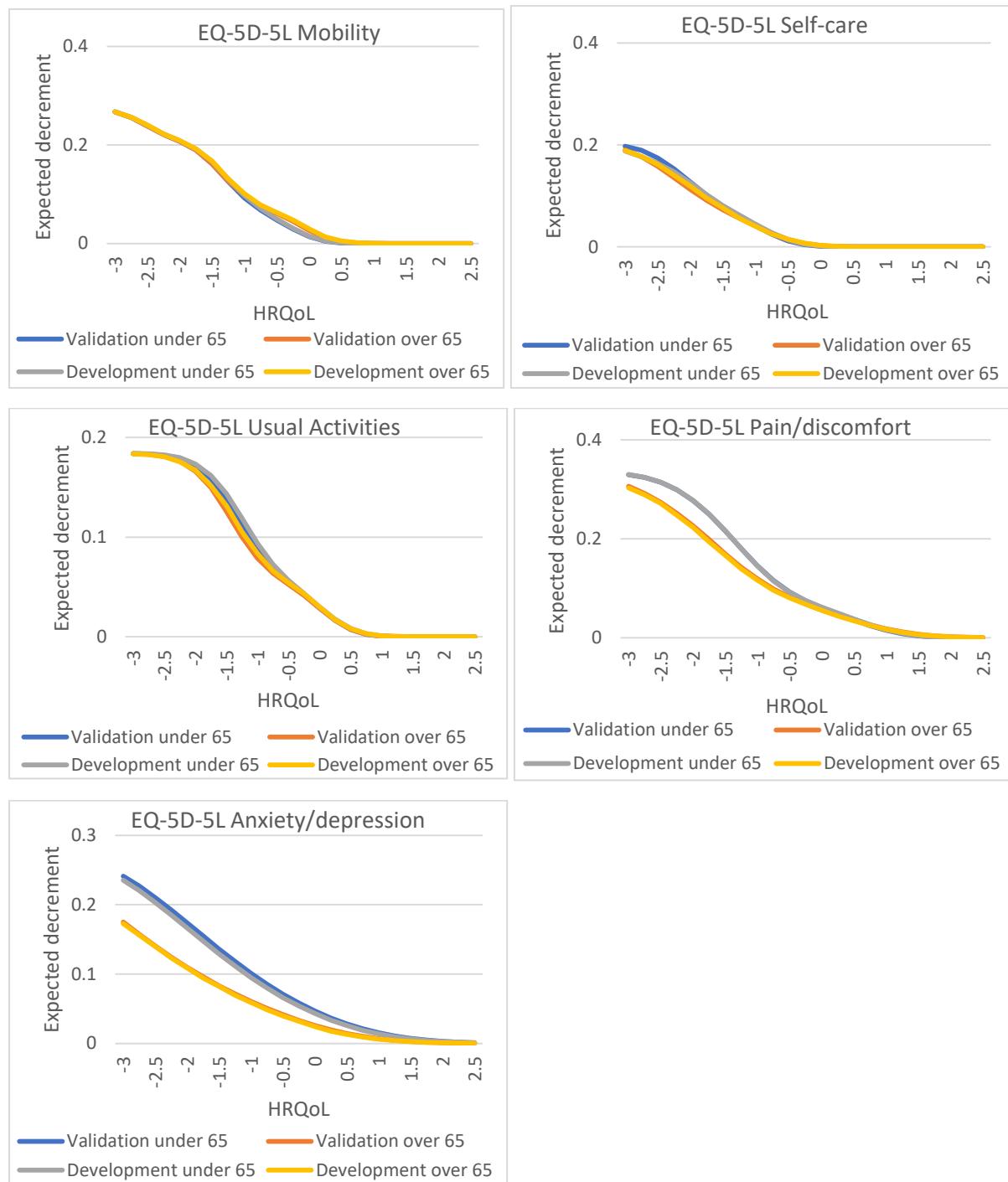
Item	Para-meter	Under 65			Over 65		
		Development	Validation	Difference	Development	Validation	Difference
Optimistic	A	0.849	0.822	0.027	0.849	0.822	0.027
	B1	-2.643	-2.682	0.039	-2.114	-2.013	-0.101
	B2	-1.570	-1.555	-0.015	-1.091	-0.931	-0.160
	B3	0.086	0.080	0.006	0.519	0.472	0.047
	B4	1.870	1.998	-0.127	1.906	1.873	0.032
Useful	A	1.127	1.144	-0.017	1.471	1.466	0.005
	B1	-2.500	-2.449	-0.050	-1.915	-1.911	-0.004
	B2	-1.737	-1.639	-0.098	-1.331	-1.279	-0.052
	B3	-0.287	-0.347	0.060	-0.020	-0.016	-0.005
	B4	1.419	1.404	0.015	1.087	1.095	-0.008
Relaxed	A	1.212	1.179	0.033	1.212	1.179	0.033
	B1	-2.537	-2.682	0.145	-2.537	-2.682	0.145
	B2	-1.273	-1.377	0.103	-1.768	-1.728	-0.040
	B3	0.263	0.238	0.025	-0.030	-0.045	0.015
	B4	1.856	1.931	-0.076	1.482	1.369	0.113
Close other people	A	0.882	0.981	-0.099	1.027	1.078	-0.051
	B1	-2.922	-2.787	-0.135	-2.509	-2.536	0.027
	B2	-1.825	-1.860	0.035	-1.868	-1.659	-0.209
	B3	-0.276	-0.346	0.070	-0.237	-0.314	0.078
	B4	1.639	1.503	0.137	1.114	1.042	0.072
Energy	A	0.858	0.945	-0.087	0.858	0.945	-0.087
	B1	-2.459	-2.197	-0.262	-1.893	-1.724	-0.169
	B2	-0.822	-0.850	0.028	-0.619	-0.514	-0.105
	B3	0.922	0.813	0.109	1.223	1.219	0.004
	B4	2.508	2.400	0.108	2.825	2.626	0.199
Deal problems	A	1.445	1.401	0.044	1.445	1.401	0.044
	B1	-2.596	-2.558	-0.038	-2.596	-2.558	-0.038
	B2	-1.825	-1.817	-0.008	-1.825	-1.817	-0.008
	B3	-0.325	-0.330	0.005	-0.325	-0.330	0.005
	B4	1.329	1.382	-0.053	1.058	1.096	-0.038
Think clearly	A	1.458	1.486	-0.028	1.458	1.486	-0.028
	B1	-2.759	-2.767	0.008	-2.759	-2.767	0.008
	B2	-1.960	-2.054	0.094	-1.960	-2.054	0.094
	B3	-0.639	-0.688	0.049	-0.639	-0.688	0.049
	B4	0.983	0.943	0.040	0.642	0.622	0.019
Feel good	A	2.075	2.083	-0.008	2.075	2.083	-0.008
	B1	-2.359	-2.305	-0.053	-2.359	-2.305	-0.053
	B2	-1.433	-1.434	0.001	-1.433	-1.434	0.001
	B3	-0.111	-0.145	0.034	0.015	0.028	-0.013
	B4	1.272	1.300	-0.028	1.125	1.185	-0.060
Close other people	A	1.357	1.293	0.064	1.357	1.293	0.064
	B1	-2.605	-2.654	0.049	-2.605	-2.654	0.049
	B2	-1.645	-1.671	0.026	-1.645	-1.671	0.026
	B3	-0.277	-0.285	0.008	-0.277	-0.285	0.008
	B4	1.214	1.279	-0.065	1.029	0.987	0.043
Confident	A	2.073	2.097	-0.024	2.073	2.097	-0.024
	B1	-2.321	-2.239	-0.082	-2.614	-2.249	-0.365
	B2	-1.422	-1.412	-0.010	-1.532	-1.470	-0.062
	B3	-0.160	-0.185	0.025	-0.072	-0.100	0.028
	B4	1.212	1.188	0.024	1.096	1.046	0.051
Make own mind	A	1.177	1.145	0.032	1.345	1.192	0.153
	B1	-3.333	-3.460	0.127	-2.917	-3.324	0.407
	B2	-2.281	-2.358	0.077	-2.314	-2.329	0.015
	B3	-0.996	-1.031	0.035	-0.871	-0.990	0.119
	B4	0.688	0.683	0.005	0.424	0.459	-0.035
Loved	A	1.122	1.07	0.052	1.122	1.07	0.052

	B1	-2.840	-2.974	0.133	-2.840	-2.974	0.133
	B2	-1.978	-2.127	0.149	-1.978	-2.127	0.149
	B3	-0.805	-0.886	0.081	-0.805	-0.886	0.081
	B4	0.397	0.453	-0.057	0.397	0.453	-0.057
Interest new things	A	1.175	1.268	-0.093	1.175	1.268	-0.093
	B1	-2.580	-2.554	-0.025	-2.300	-2.145	-0.155
	B2	-1.587	-1.522	-0.065	-1.356	-1.247	-0.109
	B3	-0.271	-0.259	-0.011	-0.047	-0.048	0.001
	B4	1.184	1.140	0.043	1.184	1.140	0.043
Cheerful	A	1.809	1.712	0.097	1.809	1.712	0.097
	B1	-2.617	-2.573	-0.044	-2.617	-2.573	-0.044
	B2	-1.671	-1.709	0.039	-1.671	-1.709	0.039
	B3	-0.339	-0.393	0.054	-0.339	-0.393	0.054
	B4	1.255	1.277	-0.022	1.088	1.040	0.048

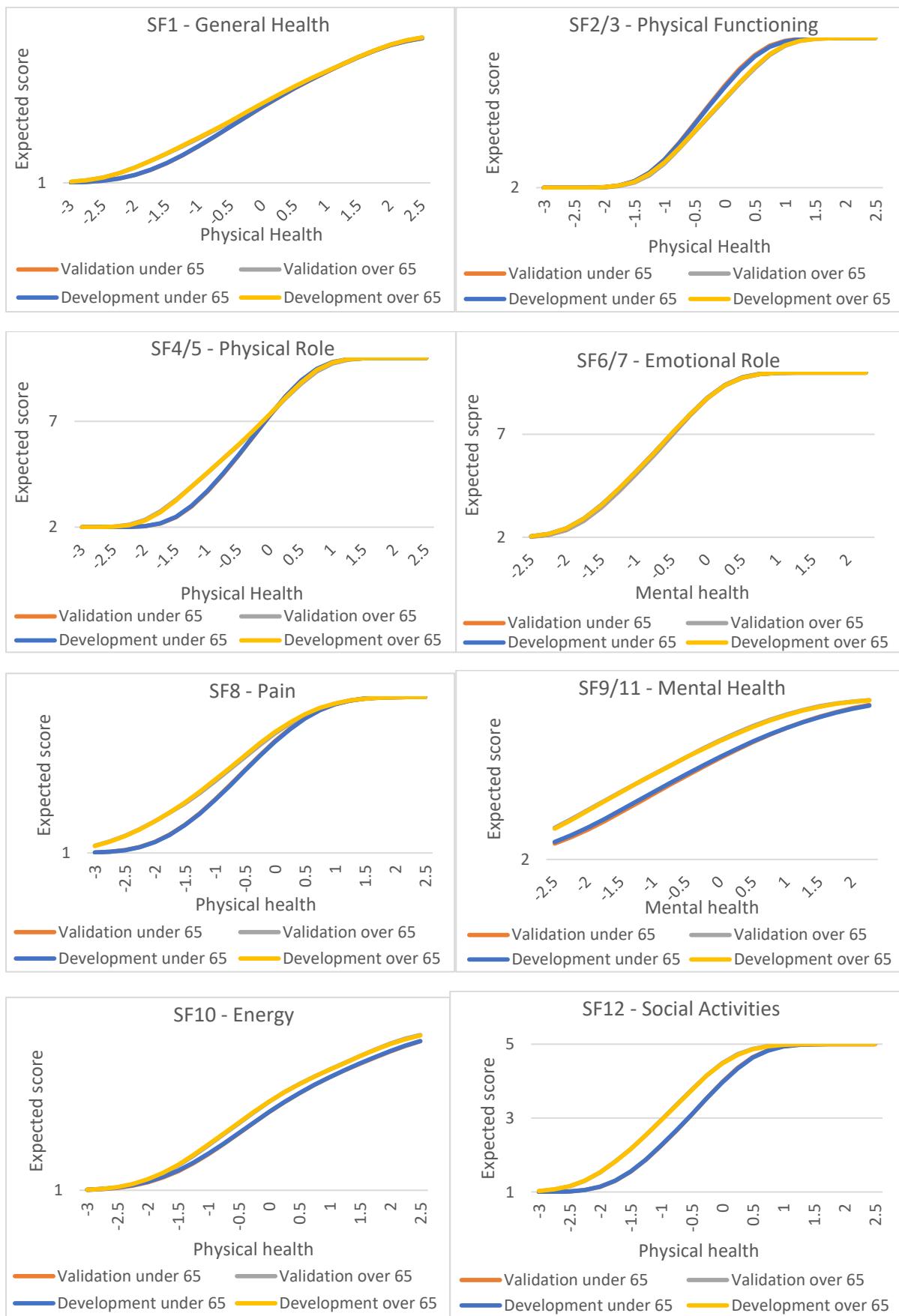
Item	Para-meter	Under 65			Over 65		
		Development	Validation	Diff	Development	Validation	Diff
Life satisfaction	a	2.511	2.58	-0.069	2.511	2.58	-0.069
	B1	-1.681	-1.648	-0.034	-1.681	-1.648	-0.034
	B2	-1.386	-1.388	0.003	-1.386	-1.388	0.003
	B3	-1.084	-1.098	0.014	-1.084	-1.098	0.014
	B4	-0.855	-0.845	-0.010	-0.855	-0.845	-0.010
	B5	-0.521	-0.497	-0.024	-0.456	-0.442	-0.014
	B6	-0.263	-0.231	-0.032	-0.263	-0.231	-0.032
	B7	0.153	0.174	-0.021	0.047	0.114	-0.066
	B8	0.810	0.763	0.047	0.570	0.688	-0.118
	B9	1.460	1.497	-0.037	1.082	1.228	-0.145
Worth-while	a	2.176	2.153	0.023	2.176	2.153	0.023
	B1	-1.843	-1.805	-0.038	-1.843	-1.805	-0.038
	B2	-1.501	-1.525	0.024	-1.501	-1.525	0.024
	B3	-1.229	-1.220	-0.009	-1.229	-1.220	-0.009
	B4	-1.015	-1.007	-0.008	-1.015	-1.007	-0.008
	B5	-0.695	-0.659	-0.036	-0.695	-0.659	-0.036
	B6	-0.488	-0.435	-0.053	-0.488	-0.435	-0.053
	B7	-0.114	-0.120	0.006	-0.234	-0.086	-0.148
	B8	0.480	0.466	0.014	0.291	0.431	-0.140
	B9	1.122	1.118	0.003	0.768	0.924	-0.156
Happy	a	3.017	2.916	0.101	3.017	2.916	0.101
	B1	-1.729	-1.654	-0.075	-1.773	-1.749	-0.024
	B2	-1.422	-1.352	-0.069	-1.446	-1.498	0.052
	B3	-1.180	-1.099	-0.081	-1.181	-1.227	0.045
	B4	-0.922	-0.859	-0.063	-0.966	-0.985	0.019
	B5	-0.623	-0.558	-0.066	-0.652	-0.611	-0.041
	B6	-0.390	-0.330	-0.060	-0.488	-0.371	-0.117
	B7	0.004	-0.005	0.009	-0.166	-0.085	-0.081
	B8	0.565	0.576	-0.011	0.274	0.392	-0.118
	B9	1.269	1.278	-0.009	0.748	0.905	-0.157
Anxious	a	1.024	1.104	-0.08	1.188	1.02	0.168
	B1	-2.341	-2.102	-0.238	-2.412	-2.410	-0.003
	B2	-1.868	-1.626	-0.242	-1.852	-1.988	0.136
	B3	-1.473	-1.250	-0.223	-1.450	-1.696	0.246
	B4	-1.203	-1.020	-0.183	-1.200	-1.423	0.222
	B5	-0.819	-0.655	-0.164	-0.795	-0.955	0.159
	B6	-0.577	-0.428	-0.149	-0.616	-0.689	0.073
	B7	-0.288	-0.175	-0.113	-0.392	-0.336	-0.056
	B8	0.091	0.134	-0.043	-0.106	-0.002	-0.104
	B9	0.494	0.482	0.012	0.216	0.293	-0.077

Appendix 20 - Expected item scores from development and validation samples

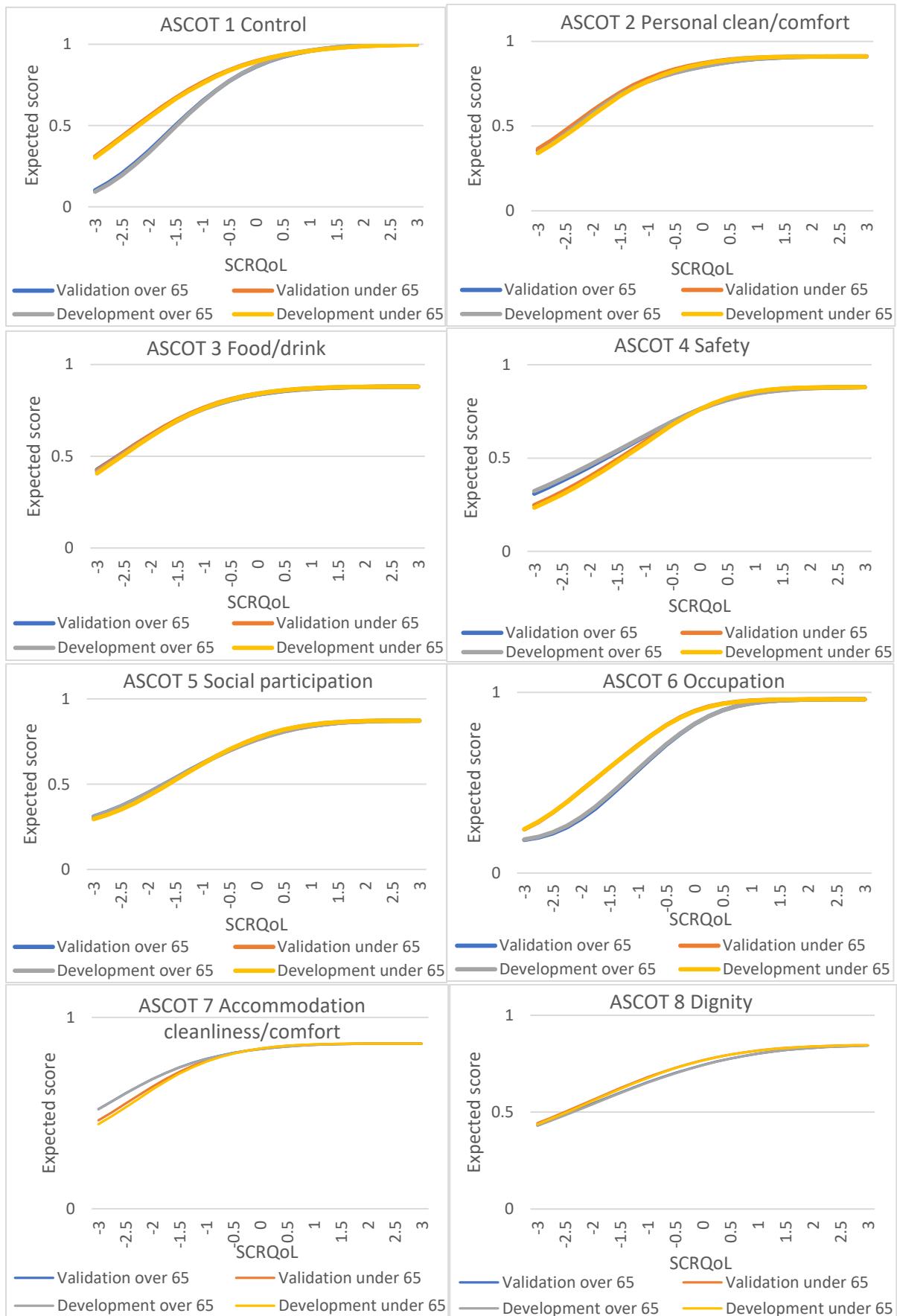
EQ-5D-5L



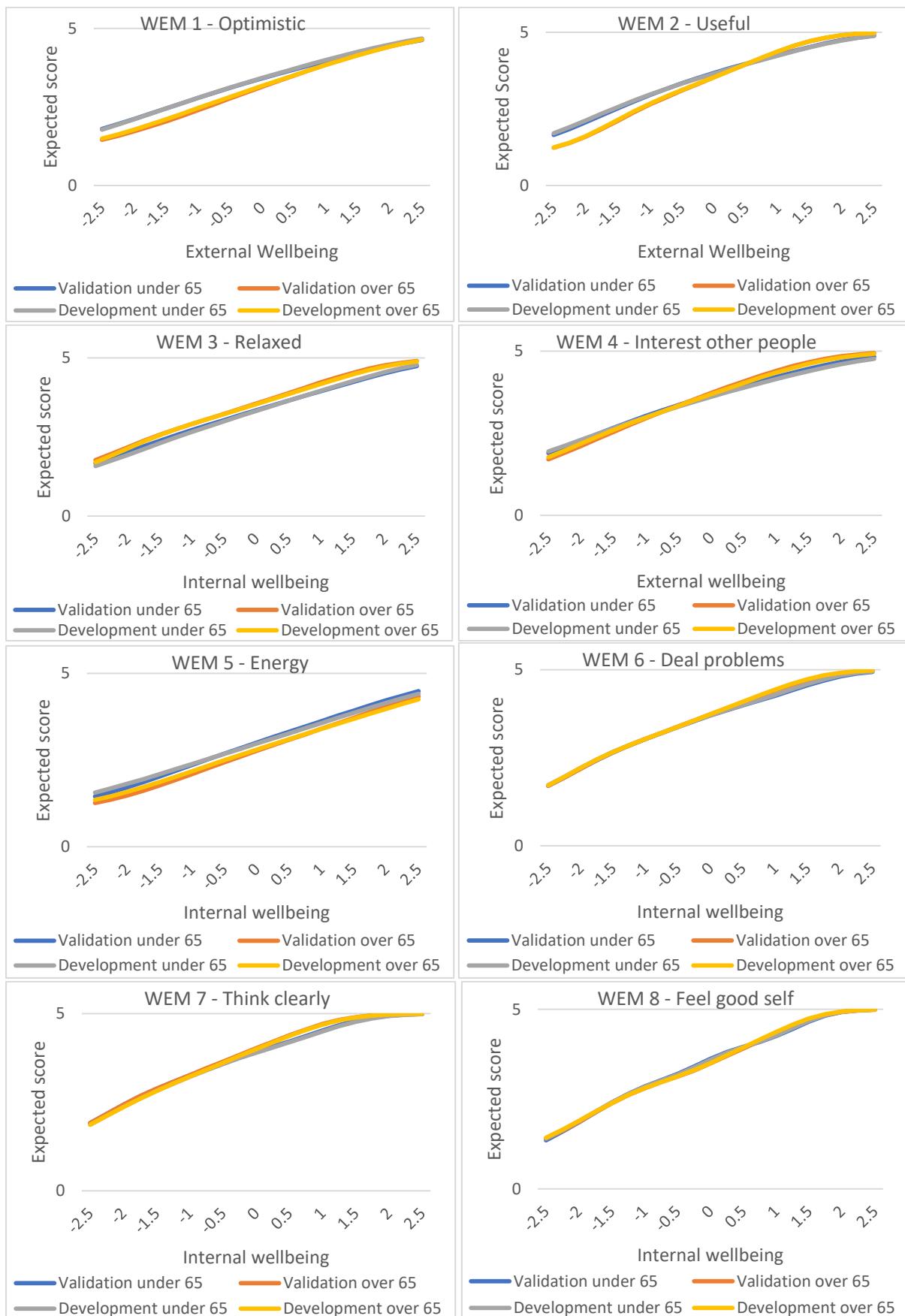
SF-12v2 TLA

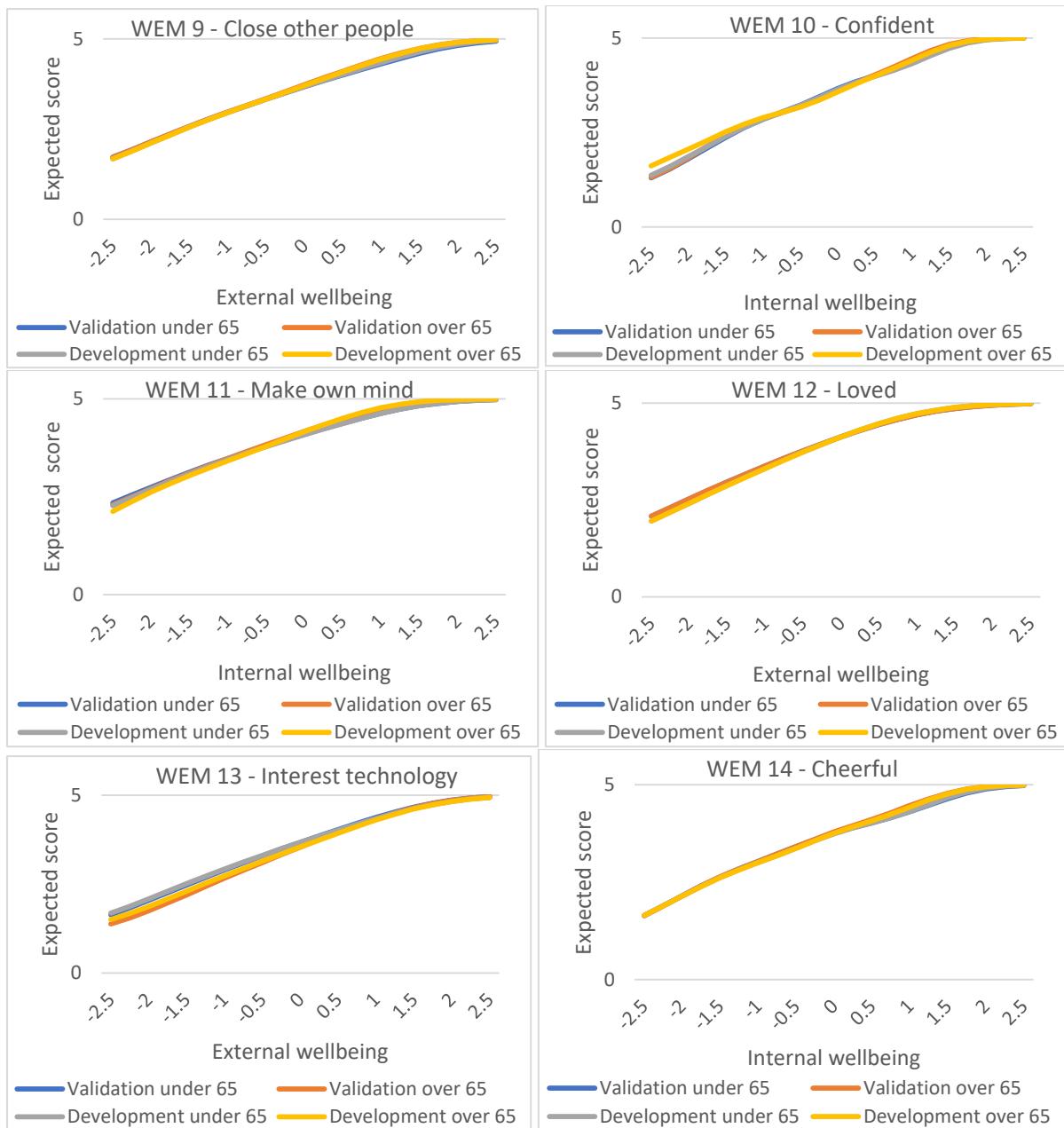


ASCOT

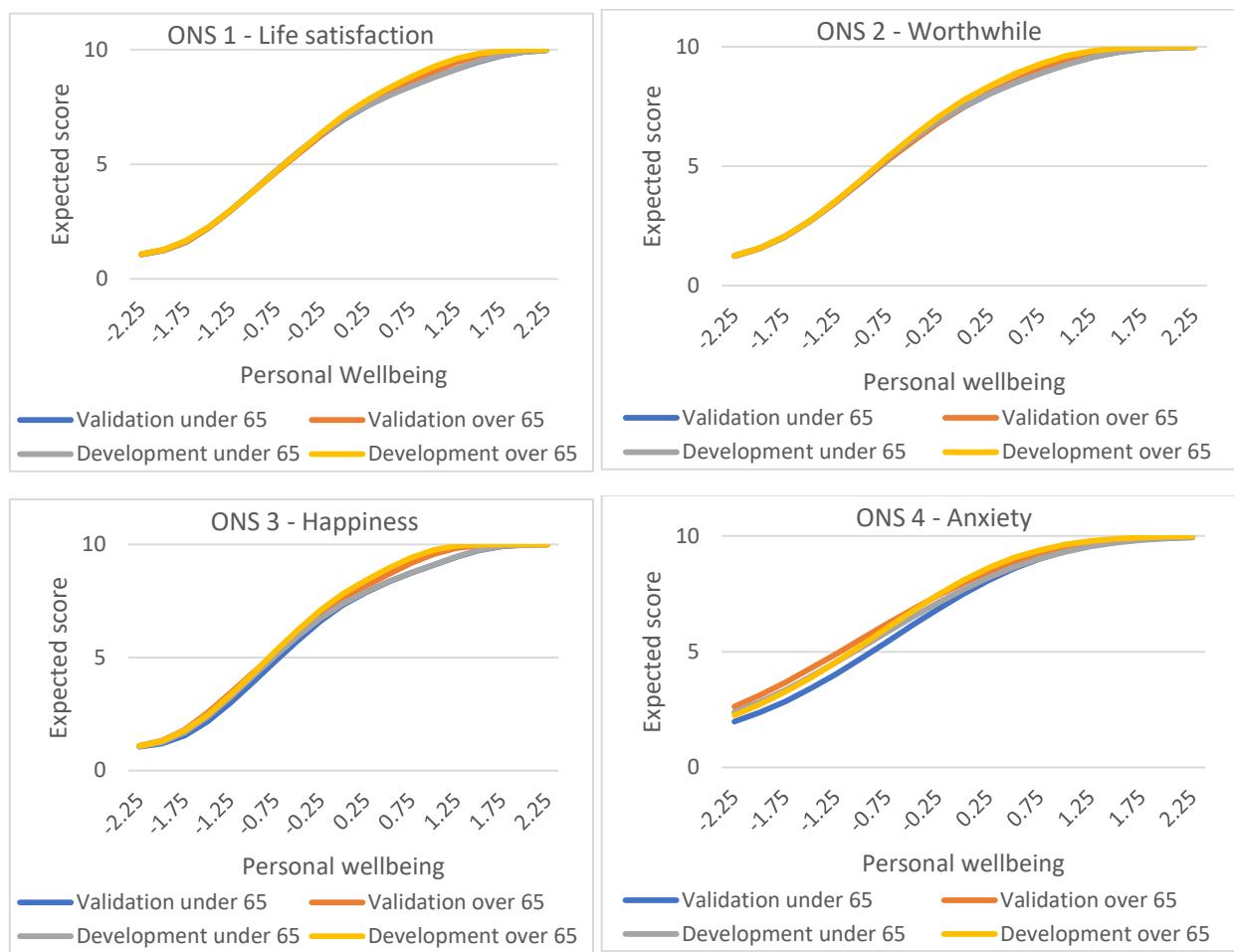


WEMWB





ONS-4



Appendix 21 – IRT sensitivity analysis - Over 75 model parameters

Under 65s					Over 65s						Under 75s					Over 75s				
a	b1	b2	b3	b4	a	b1	b2	b3	b4	EQ-5D	a	b1	b2	b3	b4	a	b1	b2	b3	b4
2.64	- 2.51	- 1.31	- 0.72	- 0.26	2.64	-2.51	-1.31	-0.52	-0.02	Mobility	2.61	- 2.48	- 1.29	- 0.65	- 0.16	2.61	- 2.48	- 1.29	- 0.48	0.00
2.47	- 2.87	- 1.94	- 1.27	- 0.76	1.86	-2.66	-2.06	-1.33	-0.86	Self-care	2.24	- 2.68	- 1.93	- 1.23	- 0.75	2.05	- 2.50	- 2.10	- 1.34	- 0.82
2.32	- 1.82	- 1.19	- 0.57	0.08	2.32	-2.03	-1.32	-0.57	0.08	Usual acts	2.34	- 1.86	- 1.19	- 0.54	- 0.11	2.34	- 1.98	- 1.38	- 0.54	0.11
1.69	- 2.18	- 1.26	- 0.42	0.61	1.30	-2.83	-1.64	-0.55	0.50	Pain/ discomf	1.59	- 2.17	- 1.25	- 0.39	- 0.64	1.23	- 3.18	- 1.80	- 0.62	0.39
0.86	- 3.21	- 2.28	- 1.26	- 0.16	0.86	-4.08	-3.26	-1.80	-0.67	Anxiety/ depress	0.85	- 1.71	- 1.27	- 0.69	- 0.09	0.85	- 3.08	- 2.40	- 1.34	- 0.54

EQ-5D-5L over 75 model parameters

SF-12v2 over 75 model parameters

Under 65s					Over 65s						Under 75s					Over 75s				
a	b1	b2	b3	b4	a	b1	b2	b3	b4	SF-12v2	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1.4	- 1.45	- 0.52	0.34	1.64	1.40	-2.14	-0.86	0.34	1.64	Gen health	1.47	-1.42	-0.48	0.47	1.70	1.46 5	- 1.81	- 0.48	0.47	1.70
2.08 3	- 0.81	0.17			2.08	-1.02	0.17			Mod Acts	2.13	-0.74	0.23			2.12 5	- 0.59	0.51		
1.60 5	- 0.77	0.13			1.61	-0.83	0.30			Stairs	1.67	-0.70	0.22			1.67 4	- 0.40	0.70		
3.85	- 1.19	- 0.59	- 0.09	0.41	2.22	-1.82	-1.02	-0.16	0.47	PR accomp	3.59	-1.17	-0.56	-0.02	0.50	3.03 5	- 1.35	- 0.60	0.15	0.67
4.20 3	- 1.09	- 0.55	- 0.10	0.42	2.29	-1.76	-1.01	-0.18	0.45	PR limited	3.73	-1.11	-0.56	-0.02	0.50	3.72 6	- 1.29	- 0.56	0.12	0.65
3.61 3	- 1.59	- 0.98	- 0.44	0.03	3.61	-1.59	-0.98	-0.44	0.03	ER accomp	4.09	-1.56	-0.96	-0.43	0.02	4.08 7	- 1.56	- 0.96	- 0.43	0.02
3.88 6	- 1.59	- 1.05	- 0.50	0.06	3.89	-1.59	-1.05	-0.50	-0.06	ER careful	4.03	-1.59	-1.04	-0.50	-0.05	4.02 9	- 1.59	- 1.04	- 0.50	- 0.05
1.62	- 1.61	- 0.80	- 0.32	0.46	1.23	-2.63	-1.40	-0.75	0.15	Pain	1.59	-1.65	-0.81	-0.31	0.48	1.27 8	- 2.26	- 1.01	- 0.38	0.32
1.08 9	- 2.04	- 1.03	- 0.05	1.84	1.09	-2.60	-1.49	-0.49	1.26	Calm/ peaceful	1.11	-2.06	-1.07	-0.12	1.66	1.10 7	- 2.52	- 1.61	- 0.60	1.02
1.21 4	- 1.31	- 0.46	0.44	1.98	1.21	-1.78	-0.90	0.16	1.80	Energy	1.27	-1.28	-0.46	0.45	2.00	1.39 5	- 1.30	- 0.42	0.41	1.82
1.07 3	- 2.28	- 1.39	- 0.34	0.72	0.90	-2.85	-2.00	-0.79	0.38	Downhearted/ low	1.00	-2.40	-1.46	-0.40	0.63	1.00 3	- 2.80	- 2.06	- 0.87	0.16
1.89 7	- 1.44	- 0.86	- 0.24	0.24	1.58	-2.20	-1.45	-0.75	-0.28	Social acts	1.87	-1.50	-0.89	-0.27	0.21	1.87 2	- 1.65	- 0.99	- 0.40	0.02

WEMWBS over 75 model parameters

Under 65s					Over 65s					WEMWBS	Under 75s					Over 75s				
a	b1	b2	b3	b4	a	b1	b2	b3	b4		a	b1	b2	b3	b4	a	b1	b2	b3	b4
0.85	-2.64	-1.57	0.09	1.87	0.85	-2.11	-1.09	0.52	1.91	Optimistic	0.82	-2.64	-1.54	0.10	1.93	0.70	-2.24	-0.96	0.88	2.22
1.13	-2.50	-1.74	-0.29	1.42	1.47	-1.92	-1.33	-0.02	1.09	Useful	1.16	-2.44	-1.71	-0.31	1.35	1.44	-1.96	-1.18	0.08	1.08
1.22	-2.54	-1.27	0.26	1.86	1.22	-2.54	-1.77	-0.03	1.48	Relaxed	1.21	-2.60	-1.38	0.17	1.76	1.21	-2.78	-1.96	-0.16	1.06
0.88	-2.92	-1.83	-0.28	1.64	1.03	-2.51	-1.87	-0.24	1.11	Interested people	0.93	-2.88	-1.90	-0.34	1.48	1.08	-2.49	-1.64	-0.29	0.95
0.86	-2.46	-0.82	0.92	2.51	0.86	-1.89	-0.62	1.22	2.83	Energy	0.90	-2.29	-0.85	0.87	2.45	0.90	-1.69	-0.46	1.24	2.62
1.45	-2.60	-1.82	-0.33	1.33	1.45	-2.60	-1.82	-0.33	1.06	Deal probs	1.45	-2.58	-1.84	-0.37	1.26	1.45	-2.58	-1.84	-0.37	0.83
1.46	-2.76	-1.96	-0.64	0.98	1.46	-2.76	-1.96	-0.64	0.64	Think clearly	1.49	-2.77	-2.02	-0.70	0.87	1.49	-2.77	-2.02	-0.70	0.37
2.01	-2.36	-1.43	-0.11	1.27	2.01	-2.36	-1.43	0.01	1.12	Feel good	2.09	-2.34	-1.45	-0.14	1.22	1.57	-2.46	-1.50	-0.07	0.91
1.36	-2.61	-1.64	-0.28	1.21	1.36	-2.61	-1.64	-0.28	1.03	Close people	1.31	-2.69	-1.70	-0.31	1.22	1.31	-2.69	-1.70	-0.31	0.83
2.08	-2.32	-1.42	-0.16	1.21	2.08	-2.61	-1.53	-0.07	1.10	Confident	2.12	-2.31	-1.46	-0.19	1.14	2.12	-2.31	-1.46	-0.19	0.80
1.18	-3.33	-2.28	-1.00	0.69	1.36	-2.92	-2.31	-0.87	0.42	Make mind	1.21	-3.30	-2.32	-1.02	0.61	1.21	-3.30	-2.32	-1.02	0.24
1.12	-2.84	-1.98	-0.80	0.40	1.12	-2.84	-1.98	-0.80	0.40	Loved	1.08	-2.99	-2.11	-0.88	0.43	1.08	-2.99	-2.11	-0.88	0.27
1.18	-2.58	-1.59	-0.27	1.18	1.18	-2.30	-1.36	-0.05	1.18	Interest things	1.19	-2.59	-1.57	-0.26	1.18	1.19	-2.23	-1.28	-0.02	1.02
1.81	-2.62	-1.67	-0.34	1.26	1.81	-2.62	-1.67	-0.34	1.09	Cheerful	1.79	-2.60	-1.71	-0.40	1.19	1.79	-2.60	-1.71	-0.40	0.77

ONS-4 over 75 model parameters

Under 65				Over 65								Under 75				Over 75			
lsat	worth	happy	anxiety	lsat	worth	happy	anxiety		life sat	worth	happy	anx	life sat	worth	happy	anx			
2.51	2.18	3.02	1.02	2.51	2.18	3.02	1.19	discrimination	2.48	2.15	2.87	1.07	2.48	2.15	2.87	1.07			
-1.68	-1.84	-1.73	-2.34	-1.68	-1.84	-1.77	-2.41	b1	-1.74	-1.90	-1.77	-	-1.74	-1.90	-1.86	-	2.57		
-1.39	-1.50	-1.42	-1.87	-1.39	-1.50	-1.45	-1.85	b2	-1.46	-1.58	-1.46	-	-1.46	-1.58	-1.59	-	2.08		
-1.08	-1.23	-1.18	-1.47	-1.08	-1.23	-1.18	-1.45	b3	-1.16	-1.29	-1.20	-	-1.16	-1.29	-1.35	-	1.70		
-0.86	-1.02	-0.92	-1.20	-0.86	-1.02	-0.97	-1.20	b4	-0.89	-1.07	-0.96	-	-0.99	-1.07	-1.11	-	1.45		
-0.52	-0.70	-0.62	-0.82	-0.46	-0.70	-0.65	-0.80	b5	-0.54	-0.74	-0.65	-	-0.54	-0.74	-0.74	-	0.99		
-0.26	-0.49	-0.39	-0.58	-0.26	-0.49	-0.49	-0.62	b6	-0.30	-0.52	-0.43	-	-0.30	-0.52	-0.53	-	0.76		
0.15	-0.11	0.00	-0.29	0.05	-0.23	-0.17	-0.39	b7	0.10	-0.19	-0.07	-	-0.01	-0.19	-0.24	-	0.48		
0.81	0.48	0.57	0.09	0.57	0.29	0.27	-0.11	b8	0.73	0.42	0.49	0.03	0.51	0.24	0.20	-	0.15		
1.46	1.12	1.27	0.49	1.08	0.77	0.75	0.22	b9	1.41	1.05	1.16	0.41	0.94	0.66	0.61	0.09			

Chapter 5 appendices

Appendix 22 – Demographic questionnaire



Attribute Questionnaire

Below are some questions to help us to find out more about you. If there are any questions which you do not feel comfortable answering, please leave those questions blank.

1. Name: _____

2. Age: _____

3. Gender

Male

Female

4. Education

What was the highest educational qualification you attained?

No qualifications GCSE HNS/HND Diploma

AS and A level Bachelor's Degree Postgraduate

5. Ethnicity

White Mixed White/Black Caribbean Mixed White/Asian

Mixed White/Asian Mixed white/Black African Other Mixed

Black African Black Caribbean Other Black Background

Asian Indian Asian Bangladeshi Asian Pakistani

Chinese Other Asian Background Other _____

6. Do you have any long-term conditions?

Yes

No

Appendix 23 - NHS Research Ethics Committee and Health Research Authority Approval Letters



Health Research Authority

South West - Frenchay Research Ethics Committee

Level 3, Block B
Whitefriars
Lewins Mead,
Bristol BS1 2NT
Email: nrescommittee.southwest-frenchay@nhs.net
Telephone: 0207 1048 045

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

05 December 2017

Miss Hannah Penton
PhD Student
University of Sheffield
ScHARR Regent Court
30 Regent Street
Sheffield
S1 4DA

Dear Miss Penton

Study title: Qualitative investigation into the validity and acceptability of the EQ-5D-5L, the SF-12, the Warwick Edinburgh Mental Wellbeing Scale (WEMWBS) and the ONS Personal Wellbeing questions (ONS-4) for measuring quality of life and wellbeing in older people experiencing frailty
REC reference: 17/SW/0230
IRAS project ID: 231380

Thank you for your submission of 23rd November 2017, responding to the Proportionate Review Sub-Committee's request for changes to the documentation for the above study.

The revised documentation has been reviewed and approved by the sub-committee.

We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this favourable opinion letter. The expectation is that this information will be published for all studies that receive an ethical opinion but should you wish to provide a substitute contact point, wish to make a request to defer, or require further information, please contact [please contact hra.studyregistration@nhs.net](mailto:hra.studyregistration@nhs.net) outlining the reasons for your request.

Under very limited circumstances (e.g. for student research which has received an unfavourable opinion), it may be possible to grant an exemption to the publication of the study.

A Research Ethics Committee established by the Health Research Authority

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA Approval (England)/ NHS permission for research is available in the Integrated Research Application System, www.hra.nhs.uk or at <http://www.rforum.nhs.uk>.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

All clinical trials (defined as the first four categories on the IRAS filter page) must be registered on a publically accessible database. This should be before the first participant is recruited but no later than 6 weeks after recruitment of the first participant.

There is no requirement to separately notify the REC but you should do so at the earliest opportunity e.g. when submitting an amendment. We will audit the registration details as part of the annual progress reporting process.

To ensure transparency in research, we strongly recommend that all research is registered but for non-clinical trials this is not currently mandatory.

If a sponsor wishes to request a deferral for study registration within the required timeframe, they should contact hra.studyregistration@nhs.net. The expectation is that all clinical trials will be registered, however, in exceptional circumstances non registration may be permissible with prior agreement from the HRA. Guidance on where to register is provided on the HRA website.



Health Research Authority

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" above).

Approved documents

The documents reviewed and approved by the Committee are:

Document	Version	Date
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [HP Certificate of Insurance]	V1.0	20 September 2017
Interview schedules or topic guides for participants [HP Interview/Focus group schedule]	V1.0	12 September 2017
IRAS Application Form [IRAS_Form_25092017]		25 September 2017
IRAS Checklist XML [Checklist_23112017]		23 November 2017
Letters of invitation to participant [HP Participant Invitation Letter]	V2.0	13 November 2017
Letters of invitation to participant [Second Invitation Letter]	V1.0	13 November 2017
Other [CV Chris Dayson Supervisor]	V1.0	26 September 2017
Other [CV Claire Hulme Supervisor]	V1.0	26 September 2017
Other [Demographic Questionnaire V1.0 23oct17 IRAS231380]	V1.0	23 October 2017
Other [Response card]	V1.0	13 November 2017
Participant consent form [HP Participant consent form]	V2.0	23 October 2017
Participant information sheet (PIS) [HP Participant Information Sheet]	V3.0	13 November 2017
Referee's report or other scientific critique report [HP Reviewer Comments]	V1.0	12 September 2017
Research protocol or project proposal [HP Protocol]	V3.0	13 November 2017
Summary CV for Chief Investigator (CI) [HP Chief Investigator CV]	V1.0	12 September 2017
Summary CV for student [HP CV Student/Chief Investigator]	V2.0	23 October 2017
Summary CV for supervisor (student research) [Tracey Young CV Supervisor]	V1.0	12 September 2017
Validated questionnaire [Validated Questionnaires]	V1.0	26 September 2017

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

A Research Ethics Committee established by the Health Research Authority

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

Feedback

You are invited to give your view of the service that you have received from the Research Ethics Service and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:

<http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance>

We are pleased to welcome researchers and R & D staff at our RES Committee members' training days – see details at <http://www.hra.nhs.uk/hra-training/>

17/SW/0230	Please quote this number on all correspondence
------------	--

With the Committee's best wishes for the success of this project.

Yours sincerely

PP wbertram.

Mrs Wendy Bertram
Vice-Chair

Email: nrescommittee.southwest-frenchay@nhs.net

Enclosures: "After ethical review – guidance for researchers"

Copy to: Dr Jennifer Burr

Mrs Jane Dennison, Bradford Teaching Hospitals NHS Foundation Trust

Miss Hannah Penton
PhD Student
University of Sheffield
ScHARR Regent Court
30 Regent Street
Sheffield
S1 4DA

Email: hra.approval@nhs.net

22 December 2017

Dear Miss Penton

Letter of HRA Approval

Study title: Qualitative investigation into the validity and acceptability of the EQ-5D-5L, the SF-12, the Warwick Edinburgh Mental Wellbeing Scale (WEMWBS) and the ONS Personal Wellbeing questions (ONS-4) for measuring quality of life and wellbeing in older people experiencing frailty

IRAS project ID: 231380
REC reference: 17/SW/0230
Sponsor University of Sheffield

I am pleased to confirm that HRA Approval has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications noted in this letter.

Participation of NHS Organisations in England

The sponsor should now provide a copy of this letter to all participating NHS organisations in England.

Appendix B provides important information for sponsors and participating NHS organisations in England for arranging and confirming capacity and capability. Please read Appendix B carefully, in particular the following sections:

- *Participating NHS organisations in England* – this clarifies the types of participating organisations in the study and whether or not all organisations will be undertaking the same activities
- *Confirmation of capacity and capability* - this confirms whether or not each type of participating NHS organisation in England is expected to give formal confirmation of capacity and capability. Where formal confirmation is not expected, the section also provides details on the time limit given to participating organisations to opt out of the study, or request additional time, before their participation is assumed.
- *Allocation of responsibilities and rights are agreed and documented (4.1 of HRA assessment criteria)* - this provides detail on the form of agreement to be used in the study to confirm capacity and capability, where applicable.

Page 1 of 8

IRAS project ID	231380
-----------------	--------

Further information on funding, HR processes, and compliance with HRA criteria and standards is also provided.

It is critical that you involve both the research management function (e.g. R&D office) supporting each organisation and the local research team (where there is one) in setting up your study. Contact details and further information about working with the research management function for each organisation can be accessed from the [HRA website](#).

Appendices

The HRA Approval letter contains the following appendices:

- A – List of documents reviewed during HRA assessment
- B – Summary of HRA assessment

After HRA Approval

The document "*After Ethical Review – guidance for sponsors and investigators*", issued with your REC favourable opinion, gives detailed guidance on reporting expectations for studies, including:

- Registration of research
- Notifying amendments
- Notifying the end of the study

The HRA website also provides guidance on these topics, and is updated in the light of changes in reporting expectations or procedures.

In addition to the guidance in the above, please note the following:

- HRA Approval applies for the duration of your REC favourable opinion, unless otherwise notified in writing by the HRA.
- Substantial amendments should be submitted directly to the Research Ethics Committee, as detailed in the *After Ethical Review* document. Non-substantial amendments should be submitted for review by the HRA using the form provided on the [HRA website](#), and emailed to hra.amendments@nhs.net.
- The HRA will categorise amendments (substantial and non-substantial) and issue confirmation of continued HRA Approval. Further details can be found on the [HRA website](#).

Scope

HRA Approval provides an approval for research involving patients or staff in NHS organisations in England.

If your study involves NHS organisations in other countries in the UK, please contact the relevant national coordinating functions for support and advice. Further information can be found through [IRAS](#).

If there are participating non-NHS organisations, local agreement should be obtained in accordance with the procedures of the local participating non-NHS organisation.

IRAS project ID	231380
-----------------	--------

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the [HRA website](#).

HRA Training

We are pleased to welcome researchers and research management staff at our training days – see details on the [HRA website](#).

Your IRAS project ID is 231380. Please quote this on all correspondence.

Yours sincerely

Maeve Ip Groot Bluemink
Assessor

Email: hra.approval@nhs.net

Copy to: *Dr Jennifer Burr, University of Sheffield – Sponsor Contact*
Mrs Jane Dennison, Bradford Teaching Hospitals NHS Foundation Trust – Lead R&D Contact

IRAS project ID	231380
-----------------	--------

Appendix A - List of Documents

The final document set assessed and approved by HRA Approval is listed below.

Document	Version	Date
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [HP Certificate of Insurance]	V1.0	20 September 2017
Interview schedules or topic guides for participants [HP Interview/Focus group schedule]	V1.0	12 September 2017
IRAS Application Form [IRAS_Form_25092017]		25 September 2017
Letters of invitation to participant [HP Participant Invitation Letter]	V2.0	13 November 2017
Letters of invitation to participant [Second Invitation Letter]	V1.0	13 November 2017
Other [CV Chris Dayson Supervisor]	V1.0	26 September 2017
Other [CV Claire Hulme Supervisor]	V1.0	26 September 2017
Other [Demographic Questionnaire V1.0 23oct17 IRAS231380]	V1.0	23 October 2017
Other [Response card]	V1.0	13 November 2017
Participant consent form [HP Participant consent form]	V2.0	23 October 2017
Participant information sheet (PIS) [HP Participant Information Sheet]	V3.0	13 November 2017
Referee's report or other scientific critique report [HP Reviewer Comments]	V1.0	12 September 2017
Research protocol or project proposal [HP Protocol]	V3.0	13 November 2017
Summary CV for Chief Investigator (CI) [HP Chief Investigator CV]	V1.0	12 September 2017
Summary CV for student [HP CV Student/Chief Investigator]	V2.0	23 October 2017
Summary CV for supervisor (student research) [Tracey Young CV Supervisor]	V1.0	12 September 2017
Validated questionnaire [Validated Questionnaires]	V1.0	26 September 2017

Appendix B - Summary of HRA Assessment

This appendix provides assurance to you, the sponsor and the NHS in England that the study, as reviewed for HRA Approval, is compliant with relevant standards. It also provides information and clarification, where appropriate, to participating NHS organisations in England to assist in assessing and arranging capacity and capability.

For information on how the sponsor should be working with participating NHS organisations in England, please refer to the, participating NHS organisations, capacity and capability and Allocation of responsibilities and rights are agreed and documented (4.1 of HRA assessment criteria) sections in this appendix.

The following person is the sponsor contact for the purpose of addressing participating organisation questions relating to the study:

Name: Dr Jennifer Burr

Tel: 01142220792

Email: j.a.burr@sheffield.ac.uk

HRA assessment criteria

Section	HRA Assessment Criteria	Compliant with Standards	Comments
1.1	IRAS application completed correctly	Yes	The Applicant confirmed that CARE 75+ cohort database is held on the Bradford Teaching Hospitals NHS Foundation Trust server and is managed by the CARE 75+ research team who work in the Academic Unit of Elderly Care & Rehabilitation based in NHS Trust premises. The CARE 75+ cohort study has REC and HRA Approval in place and is currently still running.
2.1	Participant information/consent documents and consent process	Yes	No comments
3.1	Protocol assessment	Yes	No comments
4.1	Allocation of responsibilities and rights are agreed and documented	Yes	Given the nature of the study, the Site has agreed that no exchange of contracts or SOE and SOA are needed,

Section	HRA Assessment Criteria	Compliant with Standards	Comments
			the CI should just inform the site when they are ready to begin.
4.2	Insurance/indemnity arrangements assessed	Yes	<p>Sponsor's insurance policy will cover the design, management and conduct of the study.</p> <p>Where applicable, independent contractors (e.g. General Practitioners) should ensure that the professional indemnity provided by their medical defence organisation covers the activities expected of them for this research study</p>
4.3	Financial arrangements assessed	Yes	<p>No application for external funding has been made.</p> <p>There will be no financial provisions to the sites.</p>
5.1	Compliance with the Data Protection Act and data security issues assessed	Yes	The Applicant confirmed that only anonymised data would be shared with the sponsor and funding body for audit purposes.
5.2	CTIMPS – Arrangements for compliance with the Clinical Trials Regulations assessed	Not Applicable	No comments
5.3	Compliance with any applicable laws or regulations	Yes	No comments
6.1	NHS Research Ethics Committee favourable opinion received for applicable studies	Yes	REC Favourable Opinion was issued by the South Central - Frenchay REC.
6.2	CTIMPS – Clinical Trials Authorisation (CTA) letter received	Not Applicable	No comments
6.3	Devices – MHRA notice of no objection received	Not Applicable	No comments
6.4	Other regulatory approvals and authorisations received	Not Applicable	No comments

Participating NHS Organisations in England

This provides detail on the types of participating NHS organisations in the study and a statement as to whether the activities at all organisations are the same or different.

There is one type of participating NHS organisation in England; therefore, there is only one site type.

Bradford Teaching Hospitals NHS Foundation Trust has been identified as the NHS organisation involved in this study.

If this study is subsequently extended to other NHS organisation(s) in England, an amendment should be submitted to the HRA, with a Statement of Activities and Schedule of Events for the newly participating NHS organisation(s) in England.

The Chief Investigator or sponsor should share relevant study documents with participating NHS organisations in England in order to put arrangements in place to deliver the study. The documents should be sent to both the local study team, where applicable, and the office providing the research management function at the participating organisation. For NIHR CRN Portfolio studies, the Local LCRN contact should also be copied into this correspondence. For further guidance on working with participating NHS organisations please see the HRA website.

If chief investigators, sponsors or principal investigators are asked to complete site level forms for participating NHS organisations in England which are not provided in IRAS or on the HRA website, the chief investigator, sponsor or principal investigator should notify the HRA immediately at hra.approval@nhs.net. The HRA will work with these organisations to achieve a consistent approach to information provision.

Confirmation of Capacity and Capability

This describes whether formal confirmation of capacity and capability is expected from participating NHS organisations in England.

Participating NHS organisations in England will be expected to formally confirm their capacity and capability to host this research.

- Following issue of this letter, participating NHS organisations in England may now confirm to the sponsor their capacity and capability to host this research, when ready to do so. How capacity and capability will be confirmed is detailed in the *Allocation of responsibilities and rights are agreed and documented* (4.1 of HRA assessment criteria) section of this appendix.
- The [Assessing, Arranging, and Confirming](#) document on the HRA website provides further information for the sponsor and NHS organisations on assessing, arranging and confirming capacity and capability.

Principal Investigator Suitability

This confirms whether the sponsor position on whether a PI, LC or neither should be in place is correct for each type of participating NHS organisation in England and the minimum expectations for education, training and experience that PIs should meet (where applicable).

Principal Investigators (PIs) are expected for this type of study.

GCP training is not a generic training expectation, in line with the [HRA/MHRA statement on training](#)

expectations.

HR Good Practice Resource Pack Expectations

This confirms the HR Good Practice Resource Pack expectations for the study and the pre-engagement checks that should and should not be undertaken

All research activities at sites are undertaken by members of the clinical teams employed by the host NHS organisations, therefore no additional HR arrangements (honorary research contracts or Letters of Access) will be expected.

Other Information to Aid Study Set-up

This details any other information that may be helpful to sponsors and participating NHS organisations in England to aid study set-up.

- The applicant has indicated that they do not intend to apply for inclusion on the NIHR CRN Portfolio.
- Some activity will take place outside the NHS. HRA approval does not cover activity outside the NHS. Before undertaking activity outside the NHS the research team must follow the procedures and governance arrangements of responsible organisations.

Appendix 24 – University sponsorship letter



The
University
Of
Sheffield.

School Of
Health
And
Related
Research.

ScHARR

Charlotte Claxton
Ethics Committee Administrator
Regent Court
30 Regent Street
Sheffield S1 4DA

29 July 2019

Telephone: +44 (0) 114 222 5446
Email: c.claxton@sheffield.ac.uk

Project title: PhD validity of quality of life measures in older people

Reference Number: 154182

LETTER TO CONFIRM THAT THE UNIVERSITY OF SHEFFIELD IS THE PROJECT'S RESEARCH GOVERNANCE SPONSOR

The University has reviewed the following documents:

1. A University approved costing record;
2. Confirmation of independent scientific approval;
3. Confirmation of independent ethics approval.

All the above documents are in place. Therefore, the University now **confirms** that it is the project's research governance sponsor and, as research governance sponsor, **authorises** the project to commence any non-NHS research activities. Please note that HRA approval will be required before the commencement of any activities which do involve the NHS.

You are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Good Research & Innovation Practices Policy: <https://www.sheffield.ac.uk/rs/ethicsandintegrity>, Ethics Policy: <https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/index> and Data Protection Policies: www.shef.ac.uk/cics/records. More details can be found on the University's research governance website: <https://www.sheffield.ac.uk/rs/ethicsandintegrity/governance>.

Your Supervisor, with your support and input, is responsible for providing up-to-date study documentation to all relevant sites, and for monitoring the project on an ongoing basis. Your Head of Department is responsible for independently monitoring the project as appropriate. The project may be audited during or after its lifetime by the University. The monitoring responsibilities are listed in **Annex 1**.

Yours sincerely

Charlotte Claxton

On behalf of the ScHARR Research Ethics Committee



cc. Supervisor: Kathryn Rooney

Head of Department/School: Professor John E Brazier

Monitoring responsibilities of the Supervisor:

The primary responsibility for project monitoring lies with the Supervisor. You agree to:

1. Establish a **site file** before the start of the project and ensure it remains up to date over the project's entire lifetime:
<https://www.sheffield.ac.uk/rs/ethicsandintegrity/governance/rg-forms>
 2. Provide **progress reports/written updates** to the Head of Department at reasonable points over the project's lifetime, for example at:
 - a. three months after the project has started; and
 - b. on an annual basis (only if the project lasts for over 18 months); and
 - c. at the end of the project.

See: <https://www.sheffield.ac.uk/rs/ethicsandintegrity/governance/rg-forms>
 3. Report **adverse events**, should they occur, to the Head of Department:
<https://www.sheffield.ac.uk/rs/ethicsandintegrity/governance/rg-forms>
 4. Provide progress reports to the research funder (if externally-funded).
 5. Establish appropriate arrangements for recording, reporting and reviewing significant developments as the research proceeds, and taking appropriate steps to address them – i.e. developments that have a significant impact in relation to one or more of the following:
 - the safety and well-being of the participants in the project;
 - the project's scientific direction;
 - the conduct or management of the project, including the suitability of the protocol.

The Head of Department should be alerted to significant developments in advance wherever possible.
 6. Establish appropriate arrangements to record, handle and, as appropriate, store all information collected for or as part of the research project in such a way that it can be accurately reported, interpreted and verified without compromising the confidentiality of individual care users.
 7. Establish appropriate arrangements for making information about the findings of the research accessible (and data and tissue where appropriate, with adequate consent and privacy safeguards, in a timely manner after the research has finished)
- *****

Monitoring responsibilities of the Head of Department

You agree to:

1. Review the **standard monitoring progress reports**, submitted by the Supervisor, and follow up any issues or concerns that the reports raise with the Supervisor.
2. Verify that **adverse events**, should they occur, have been reported properly and that actions have been taken to address the impact of the adverse event(s) and/or to limit the risk of similar adverse event(s) reoccurring.
3. Verify that a project is complying with any **ethics conditions** (e.g. that the information sheet and consent form approved by ethics reviewers is being used; e.g. that informed consent has been obtained from participants).
4. Introduce a form of **correspondence** (e.g. regular email, annual meeting) with a project's Supervisor, that is **proportionate to the project's potential level of risk**, in order to verify that a project is complying with the approved protocol and/or with any research funder conditions. Whatever correspondence is chosen the Head of Department should, as a minimum, ensure that s/he is informed sufficiently in advance about significant developments wherever possible.

Appendix 25 – Participant information sheet

Participant information sheet



1. Research Project Title:

The validity and acceptability of quality of life and wellbeing measures in older people

2. Invitation

You are invited to take part in a research project investigating whether currently used quality of life and wellbeing questionnaires are appropriate and relevant to older people. Before you decide whether or not you wish to take part, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Please contact us if anything is not clear or if you would like more information. Our contact details are at the end of this information sheet.

3. What is the project's purpose?

I am a PhD student from the School of Health and Related Research at the University of Sheffield. This research focuses on the measurement of quality of life and wellbeing in older people. Questionnaires measuring quality of life and wellbeing are commonly used to measure the benefits of new and existing health care treatments and services. This information is used to decide which health and care services to provide through the NHS and social services.

The validity of these questionnaires, in terms of how relevant the questions are to people's quality of life and how appropriate the questions are to ask, has been tested in younger groups but not in older people. This research aims to investigate how relevant and appropriate currently used quality of life and wellbeing questionnaires are to older people to be sure that accurate estimates of quality of life and wellbeing are made and the best services for older people are funded and provided by the NHS.

4. Why have I been chosen?

You have been approached to take part because you are an older person who, as a participant in the CARE 75+ cohort study, has experience in being asked questions about your health and quality of life.

5. Do I have to take part?

You do not have to take part in this study. Participation is voluntary and the decision is completely up to you. You are free to withdraw at any point without giving reason for doing

Participant information sheet



so. Your decision whether or not to take part has no impact on your existing role in the CARE 75+ study. There will be no negative consequences if you decide not to participate.

6. What do I have to do and what will happen to me if I take part?

If you choose to take part in the study you will participate in either an individual interview or a group interview; the choice is yours. Before the individual or group interview begins you will be given a consent form to sign to say you are happy to undertake the interview.

Both the individual and group interviews will cover the same topic and ask the same questions. You will be asked to fill out questionnaires which are commonly used to assess people's quality of life and wellbeing and then you will be asked your opinions about the questionnaires and the questions included in them. Sessions will last up to 2 hours, with breaks provided. You will only be asked to attend one session, not both.

Individual interviews will be carried out with just you and the researcher (Hannah Penton) in your own home. Group interviews will be carried out at convenient, accessible locations, with you, the researcher and up to 6 other participants present. In group interviews you will all sit together. The researcher will ask questions and you will be able to discuss your thoughts and opinions as a group. In individual interviews, the researcher will ask you questions and you can answer privately.

If you choose to attend a group interview, you will have to travel to the agreed location of the group interview. Group interview attendees will be offered up to a maximum of £10 reimbursement for travel costs. If you choose an individual interview this will be carried out in your home. We will not be able to reimburse you as you will not be required to travel.

7. Will I be recorded and how will the recording be used?

The interview will be audio recorded and transcribed by the researcher (Hannah Penton). Transcripts will be anonymous, with all personal information removed. Nobody outside the research team (Hannah Penton and her 3 supervisors) will have access to the recording. The recording will be destroyed after the project is finished (in approximately 1 year). All audio recordings and transcripts will be stored in a safe secure location on the University of Sheffield campus. This study has to be audio recorded for the researcher to be able to accurately and fully analyse and report the views of participants. Unfortunately, the study cannot accept participants who do not want to be audio recorded.

8. What are the possible disadvantages and risks of taking part?

The interview will discuss how relevant and acceptable questions about different aspects of quality of life and wellbeing are to ask older people. If you find the discussion of quality of life and wellbeing distressing, you are not required to participate in the study.

If at any stage during an individual interview you are distressed or uncomfortable about topics being asked you can: ask the researcher to move on, ask the researcher to pause the interview or ask the researcher to stop the interview. If you become distressed or uncomfortable during the group interview and you want to leave you have the right to leave at any time. If you need a break during the group interview, or you find a particular topic uncomfortable, you are welcome to leave and return as you wish. You can also choose not to answer some questions if you are uncomfortable with a particular topic.

If you find the interview process excessively tiring and feel you need a break you can make this known to the researcher at any stage. Individual interviews are easier to pause immediately and for as long as necessary so if you feel you are likely to become exhausted quickly, the individual interview may be the best option for you.

9. What are the possible benefits of taking part?

By taking part in this research you will contribute to our understanding of how well existing measures of quality of life and wellbeing perform in older people. These measures are used to measure the benefits of health and social services to older people. Decisions on which NHS and social services to provide to people are based on these measured benefits. We hope this research will help us better understand which measures provide the most accurate estimate of the quality of life and wellbeing of older people.

10. What happens if the research study stops earlier than expected?

If for any reason the research has to be stopped earlier than expected, you will be contacted immediately and informed. If you do not meet the eligibility criteria for this study, then you will be informed as soon as possible.

11. What if something goes wrong?

If something goes wrong or you would like to raise a complaint during the study you can contact either Miss Hannah Penton (email: hpenton1@sheffield.ac.uk telephone: [07843840476](tel:07843840476)) or Dr Tracey Young (email: t.a.young@sheffield.ac.uk telephone: 0114 222 0837). If for any reason you are not satisfied with the outcome you can contact Professor John Brazier, Dean of School of Health and Related Research (email:

Participant information sheet



j.e.brazier@sheffield.ac.uk; telephone: 0114 222 0726; Address: School of Health and Related Research, Regent Court, 30 Regent Street, Sheffield, S1 4DA).

12. Will my taking part in this project be kept confidential?

The information that is collected about you during this research will be kept strictly confidential. You will not be identifiable in any reports or publications. All audio recordings and transcripts are stored in a locked filing cabinet on university premises.

13. What will happen to the results of the research project?

The results of this project will form part of the researcher's PhD thesis. Data from this research will be published in scientific journals and presented at academic conferences. This will include anonymised quotes from the interviews. All results that are made public will be anonymized by removing any words that could identify you. You will not be identifiable in any publications. Anonymised data will be kept for a maximum of 5 years.

14. Who is organising and funding the research?

This research is funded jointly by the White Rose Network and the Collaboration of Leadership in Applied Health Research and Care (CLAHRC) Yorkshire and Humber, which was also involved in funding the CARE 75+ study in which you are a participant.

15. Who has ethically reviewed the project?

This project has been ethically approved by the South West – Frenchay NHS Research Ethics Committee. The application number is: 231380.

16. Contact for further information

Thank you for taking the time to read this information sheet.

If you are interested in taking part in this study, or if you would like more information, please contact Hannah Penton on:

email: hpenton1@sheffield.ac.uk telephone: 07843 840476



Or you can complete and return the response card in the included envelope and Hannah Penton will call you to discuss the project further or to arrange an individual or group interview.

Appendix 26 – Consent form

Consent Form – Participant Copy



The
University
Of
Sheffield.

Title of Project: Validity and Acceptability of Quality of Life and Wellbeing Measures in Older People

Name of Researcher: Hannah Penton

Please initial box

1. I confirm that I have read the information sheet dated explaining
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, and without any negative consequences. I understand that should I not wish to answer any particular questions I am free to decline.
3. I understand that the information collected about me will be kept strictly confidential and kept anonymous except in the case of safeguarding. I give permission for other members of the research team to have access to my anonymised responses. I understand that my name will be removed and I will not be identified or identifiable in any reports that result from the research.
4. I agree to anonymised quotes being included in any written reports and presentations.
5. I agree to the individual/group interview being audio recorded
6. I agree to take part in the above research project.
7. I agree for the anonymised data collected in this study to be used in future research (optional). Yes No

Name of participant

Date

Signature

Name of person taking consent

Date

Signature

Contact Information: Hannah Penton

Email: henton1@sheffield.ac.uk Telephone: 07843 840476

V2.0 23/10/17

IRAS 231380

Appendix 27 – EQ-5D concept guide

The following definitions of concepts were taken from Brooks, Rabin and de Charro 2003, Appendix 7 – Definitions of EQ-5D concepts (Brooks, Rabin et al., 2003)

Concept	Definition
Health	A general term relating to physical, emotional and social functioning; is wider than a strict medical interpretation (e.g. absence of illness), as it also includes emotional and social well-being. Includes both negative aspects of health (illness) as well as positive aspects (well-being)
Today	The day of completing the questionnaire (this particular calendar day)
Mobility	This refers to the physical ability to walk or move about, both inside and outside. It does not refer to the use of a bicycle, car or public transport
Walking about (mobility)	The ability to walk or move about independently from one place to another, both inside and outside. It does not refer to walking about an object, such as a building. “Walking about” does not refer to running strenuous activities, country walks or sport.
Confined to bed (mobility)	Restricted to staying in bed (except to use the toilet). It includes being confined to a chair (but not a wheelchair) all day (e.g. where someone is moved from bed to a chair and returned to bed at the end of the day). This can be a long-term condition or short term (e.g. in bed because of influenza). What is important is that the subject is confined to bed on the day the EQ-5D is administered.
Suggested interpretation of the 3 levels of mobility	Level 1: can walk about without help or aids Level 2: Needs to use stick, crutches, walking frame, when walking. Would include people in a wheelchair (although they may not classify themselves in level 2) Level 3: confined to bed or chair all of the day (except to use the toilet). Excludes people in a wheelchair.
Self-care	The term self-care refers to independence in daily personal care. It specifically covers washing and dressing, but also includes feeding oneself, personal hygiene, brushing teeth, grooming and going to the toilet. It does not include social or role activities, or the ability to manage personal finances or household affairs.
Usual Activities	This refers to activities such as work (paid and unpaid), study, housework, leisure and social activities. “usual” means activities carried out on a regular basis, but not necessarily on a daily basis. The activities should be “usual for you”, i.e. the respondent personally. The ability to perform usual activities refers to the ability to be able to participate in these activities today rather than to accomplish or complete them.

Pain	Physical or bodily hurt. Does not refer to psychological or mental suffering
Discomfort	Uncomfortable physical sensation, of a lower grade of intensity than pain. Includes aches, breathlessness, itching, palpitations, nausea, tiredness, dizziness, bloatedness, pins and needles, ringing in the ears. Does not include psychological or mental disturbance.
Anxiety	Psychological sensation related to “worry”; covers general feelings of feeling tense, troubled nervous, apprehensive, fearful. An example of extreme anxiety may be panic or dread.
Depression	Psychological sensation relating to lowness of spirit. Does not refer only to clinical depression; covers feeling cheerless, gloomy, dejected, down, sad, miserable, unhappy. No inherent time element i.e. not defined by length of time for which it has been experienced.
Some/moderate problems	Ranges from a small number or a small degree of difficulty to many problems or difficulties. Should indicate a middle level between no problems and extreme problems. More severe than mild.
Extreme	Indicating a very severe or very bad level – the highest (outermost) level.
Best imaginable health state	The most optimal, desirable, ideal health state a person can imagine.
Worst imaginable health state	The most bad, undesirable health state a person can imagine. Unable to function independently in all areas of life.

Appendix 28 – WEMWBS partial concept guide

Concept	Definition
Optimistic	An expectation that the future will be good rather than hoping it will be.
Useful	Useful to other people – the feeling that you are effective or making a contribution to your community or family.
Relaxed	
Interested in other people	
Energy to spare	Energy “to spare” just means plenty of energy (NOT more than usual)
Dealing with problems well	“Well” refers to the present time. Please don’t translate it as “better” or “extra”.
Thinking clearly	
Feeling good about myself	
Close to other people	“Close” not closer –same reason as above.
Confident	
Able to make up own mind about things	This question is about being capable of making decisions or having opinions.
Loved	
Interested in new things	This implies new activities and interests.
Cheerful	

Reproduced from *Frequent issues in translation*. Found at

https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs/researchers/languages/frequent_issues_in_translation.pdf Accessed 8th October 2018