CAPABILITY AS AN OUTCOME MEASURE IN RANDOMISED CONTROLLED TRIALS

by

THOMAS JAMES HIER KEELEY

A thesis submitted to the University of Birmingham for the degree of Doctor of Philosophy.

MRC Midlands Hub for Trials Methodology Research

School of Health and Population Sciences

College of Medicine and Dentistry

The University of Birmingham

January 2014

Abstract

‘The capability approach is a broad, normative framework for the evaluation of well-being’(p.94)[1], which has attracted growing interest in health and health economics research. A broader measure of well-being may more accurately capture the effects of some interventions, than traditional health-related quality of life measures. The ICECAP-A and ICECAP-O are two measures of a person’s well-being, with a theoretical grounding in the capability approach, designed for use in health and social care research.

This thesis reports qualitative and quantitative investigations into the validity and responsiveness of the ICECAP measures. A methodological review of existing validation studies was completed. Seventeen semi-structured interviews with health research professionals were carried out and an iterative, constant comparative, thematic analysis was completed to assess the content validity of the ICECAP-A. The construct validity and responsiveness of the measures were assessed using two randomised controlled trials: the BEEP trial (ISRCTN 93634563) and the Past BP trial (ISRCTN 29062286).

Qualitative and quantitative results provide positive indications of validity. The qualitative work showed that research professionals viewed the ICECAP-A as a relevant and feasible measure for use in health research. The quantitative results confirmed the majority of *a priori* hypotheses in the validity analyses, while longitudinal data provided evidence that the measures are responsive to self-reported changes in health status.

In conclusion, this thesis reports the first assessment of validity in a randomised controlled trial setting and the first analysis of responsiveness. While further testing of the ICECAP

measures is required, results indicate that the measures are appropriate for use in health research.

Dedication

For Margaret Hier, who loved the idea of this work and always wanted to know if I was "getting any good results". Well, here they are Grandma.

Acknowledgements

To my supervisors, Prof. Joanna Coast and Dr Hareth Al-Janabi for their continued and enthusiastic support for the duration of this work. The work has been substantially improved, and my learning experienced greatly enhanced, through their direction and input.

To members of the Midlands Hub for Trials Methodology Research for their continued support. Special thanks to Prof. Cindy Billingham for her time and encouragement and Karen Biddle for her invaluable support on a whole range of things.

To the members of the PastBP trial research team at the University of Birmingham and the BEEP trial research team at Keeley University, for the timely provision of data for this analysis. Special thanks to Kate Fletcher at the University of Birmingham and Elaine Nicholls at Keele University.

To Dr Paula Lorgelly at Monash University, Melbourne, for helping make the overseas research trip a reality and for her guidance while in Australia.

To Dr Laura Pearson for proof reading the entire thesis and being there throughout.

This PhD studentship was funded by the Medical Research Council Midland Hub for Trials Methodology Research. The Medical Research Council Grant ID number is G0800808.

Contents

List of tables

List of figures and illustrations

List of Abbreviations

ICECAP-A	ICEpop CAPability measure for Adults
ICECAP-O	ICEpop CAPability measure for Older people
BEEP trial	Improving the effectiveness of exercise for knee pain in older adults in primary care: Benefits of Effective Exercise for knee Pain (BEEP)
PastBP trial	A randomised controlled trial of different blood pressure targets for people with a history of stroke or transient ischaemic attach (TIA) in primary care.
EQ-5D	EuroQol 5 Dimension
EQ-5D VAS	EuroQol 5 Dimension Visual Analogue Scale
GAD-7	Generalised Anxiety Disorder assessment
HUI	Health Utilities Index
PHQ-8	Patient Health Questionnaire depression scale
SF-36	Short Form (36) health survey
WOMAC	Western Ontario and McMaster Universities Arthritis Index

Introduction

The measurement of quality of life is of importance in randomised controlled trials and in health economic evaluations occurring alongside these trials. Quality of life measurement has been dominated by a focus on health-related quality of life, with only a minority of existing measures focusing on the broader aspects of quality of life. The capability approach is a theoretical basis upon which a broader measurement of quality of life can be based. Efforts to apply the capability approach to quality of life measurement in the area of health research have been undertaken in recent years. The question of whether capability based measures accurately capture the construct which they are designed to measure (capability) is of utmost importance in the design, testing and use of these measures. This thesis explores the validity and responsiveness of two capability based quality of life measures being used in health research: the ICECAP-O and ICECAP-A measures.

Chapter one describes the evolution in normative health economics from welfarism to capability. It begins by describing the key tenets of welfarism and provides a description of the limitations of this approach. The description of approaches that have gone beyond welfarism begins with extra-welfarism. The theoretical basis of this approach is provided and its practical implementation is described. Finally the theoretical basis of the capability approach is described.

Chapter two has two parts. Firstly the randomised control trial is described and the current norms of quality of life measurement in trials are considered. The use of preference based measures, often used in health economic analysis, is covered in detail. Attempts to "operationalise" the capability approach are examined, with a particular focus on the ICECAP capability measures. The second part of chapter two provides a methodological review of

studies which have sought to validate the ICECAP measures or provide validity information through their research. The current evidence base is documented and areas for future research are described.

Chapter three introduces the theory of psychometric testing. Psychometrics is an area in which the underlying theory has undergone considerable change in the last 60 years. The field is still evolving. The section will describe the evolution of psychometrics from criterion based testing to the more scientific approach of construct validation. The importance of validating the content of the descriptive system of the measure independent from the scores generated is discussed. Finally, the challenges of validating a capability measure in a randomised controlled trial are discussed.

Chapters four and six describe the qualitative and quantitative methods employed in this thesis, respectively. Chapter four describes the recruitment and interviewing of informants and the process of transcribing and analysing the data. The comparative direct approach is a novel method of assessing content validity. This was developed based on experiences of this qualitative work and is described in detail. Chapter six describes the process for recruiting trials to participate in this work as well as the data provided by the two trials which were used in the final analysis. The methods and statistical analyses used for both the validity and responsiveness analysis are described.

Chapter five presents the results of the qualitative analysis. How informants defined quality of life is examined in detail and the informants' opinions of the content of the ICECAP-A measure are described. The comparative direct approach is used to draw comparison between these. Finally, a conceptual model of how quality of life measures are selected for use in randomised controlled trials, which was an emergent theme from this work, is discussed.

Chapter seven presents the results of the construct validity analysis. Results from the PastBP and BEEP trials are presented separately. A number of different comparator measures are used. *A priori* hypotheses presented in chapter six are tested using suitable statistical analyses. Factor analysis is used to assess the relationship between items of the ICECAP measures and EQ-5D-3L.

The results of the responsiveness analysis are provided in Chapter eight. The results of the ICECAP-O analysis using data from the PastBP trial are presented first, followed by the results of the ICECAP-A analysis using the BEEP data. Contained within this chapter are descriptions of the choice of anchors and the creation of anchor groups. Numerous anchor analyses are provided for each measure.

The discussion is presented in Chapter 9. The primary contributions of the work and its strengths and weakness are discussed. The findings of both the qualitative and quantitative research are discussed through the use of broad themes of research. Directions for future research are proposed.

CHAPTER 1. WELFARISM, EXTRA-WELFARISM AND THE CAPABILITY APPROACH

Note on chapter: Small sections of this chapter have been submitted to the University of Birmingham as part of an essay towards an MSc in Health Economics and Health Policy. In line with University of Birmingham regulations, where appropriate, this essay has been referenced to indicate the use of previously submitted material.

1.1. Chapter Introduction

A primary challenge of health care financing and management is the allocation of scarce resources between many needs and wants [2–5]. Health care policy makers, responsible for priority setting, are faced with ever increasing numbers of treatments, ever increasing costs and constantly changing rates and loci of illness. Economists and health economists have sought to assist policy makers through applying the discipline of economics to health policy. Positive economics seeks to explain 'the effects of an event on objectively measureable economic variables'$_{(p.1)}$[6], while normative economics seeks to inform decisions about how resources should be allocated through economic evaluations of costs and benefits [6].

Within the field of normative economics, a spread of ideas has resulted in differing approaches and competing schools of thought to the challenge of resource allocation in health [5]. These different approaches are not clear cut and well defined. Rather, the evolution from welfarism towards extra-welfarism and the capability approach has resulted in a blurring of the boundaries between the three main approaches to health resource allocation, with common themes running through all [7].

This spread of ideas and evolution of theory has led to differences in the methods through which each approach measures outcomes and what outcomes are measured. Therefore, along with this evolution of theory, there has been a change in the outcome measures used to inform normative decisions in health care.

This chapter provides an introduction to the different philosophies of normative economic evaluation and resource allocation in health. It sequentially defines welfarism and extra-welfarism and describes the methods through which each approach measures outcomes. Critiques of each approach are presented, with a focus on the justification and reasoning for

the development of the capability approach. The main tenets of Amartya Sen's capability approach are described along with a discussion of the challenges of "operationalising" the approach.

1.2. The Welfarist approach

The following section describes neo-classical welfare economics [8]. The evolution of the utility principle is discussed and its consequentialist and individual nature is critiqued. The requirement for maximisation of an outcome, through the Pareto Criterion, is presented. This section concludes by detailing the limited nature of welfare economics as an evaluative tool and draws on Brouwer's [9] 'seeds' of extra-welfarism, in order to demonstrate the value of a different framework.

1.2.1. Theoretical basis of welfarism

'Welfare economics is the framework within which the normative significance of economic events is evaluated'$_{(p.1)}$[10]. It attempts to assess the consequences and value of an event for economic arrangements. In health this framework can be used for the economic evaluation of treatments and technologies. While the ultimate goal of welfare economics (that of using the instruments of economics for the improvement of human life [11]) has not changed, the framework within which this should be done has been under constant evolution. The classical interpretation, proposed in Arthur Pigou's landmark text *The Economics of Welfare* [11], takes welfare to be an interpersonally comparable commodity that can be summed across individuals, allowing a sum maximum to be reached. However, as a result of convincing critiques by authors such as Lionel Robbins [12] and others, which highlight the impossibility of making objective interpersonal comparisons of welfare, a neo-classical form of welfare economics has emerged. This school of thought has, for now, secured welfare economics in 'the informational basis of ordinal and interpersonally non-comparable welfare'$_{(p.3)}$[13].

1.2.1.1. Utility in welfare economics

The attitude within economics towards utility has always been ambiguous and the meaning of utility has changed over time [14]. The current characterisation, although widely accepted in welfare economics, is still the focus of debate and empirical research. The implications of these different meanings are important, and the careless use of this term can cause confusion.

The original interpretation, which was proposed by Jeremy Bentham, forwarded utility as a measure of pleasure. Bentham [15,16] suggested that utility referred to the "sovereign masters" of pleasure and pain, that ultimately determine our actions. This hedonistic interpretation of utility has been termed *experienced utility* [17]. Edgeworth [16,18] extended this interpretation with the suggestion of an imaginary measurement instrument: the 'hedonimeter'. Similar to a barometer, this instrument would measure positive or negative pleasure at any moment, and plot experienced utility as a function of time [19], thereby allowing the quantification of pleasure over a given period. The idea that utility was measurable on a cardinal scale was a key concept of early interpretations and allowed for interpersonal comparisons of utility.

This interpretation was, however, largely rejected by economists at the start of the twentieth century [20]. Pareto [21] stated that the cardinal measurement of utility was not required, while Robbins [12] argued that cardinal utility was not a measurable concept. This difficulty with interpersonal comparisons is aptly summed up by Hicks [22]:

> *'You cannot take a temperature when you have to use, not one thermometer, but an immense number of different thermometers, working on different principles, and with no necessary correlation between their registrations'* (p.699).

The measurement of observable choices was preferred on the basis that individuals act as rational agents and will choose the option that yields most utility. This step simplified the

measurement of utility by requiring solely the observation and the ordinal ranking of an individual's choices, and not the measurement of pleasure or pain, that results from that choice. This interpretation of utility is therefore one of human wants, intuitively defined by Irving Fisher as "wantability" [23]. This *decision utility* gives a representation of an individual's preference ordering over goods bundles or states of the world [9].

It is self-evident that if individuals are able to accurately want that which they will in turn enjoy, known as affective forecasting [19], then experienced and decision utility will yield the same outcome [24]. However, if individuals are not able to make accurate decisions about future utility, then a clear divergence in the two concepts can be seen. Empirical research suggests that this may be the case. Sieff et al. [25] found lower levels of distress in patients who had just received a positive HIV test result than was expected by the same patients before the test result was received. Redelmeier & Kahneman [26] found that patients "misremembered" their self-reported level of pain during a colonoscopy procedure, when asked about it after surgery. Furthermore, it is persuasively suggested that the *ex-ante* concept of decision utilities is affected by individual's attitudes, with a focus on transitions, while the *ex-post* concept of experienced utility is focused on experiences and states [19,20,24]. For example: the *ex-ante* attitude to being diagnosed with HIV might be one of horror, while the *ex-post* reality of living with HIV in the 21st Century is, for most, that of a manageable, chronic disease. Empirical findings and theoretical propositions such as these have led some [19,20] to re-evaluate the potential for the use of experienced utility on the basis that decision utility may lead to the systematic over or under estimation of utility. While this research is both of interest and importance, it is at an early stage. In line with the majority of modern welfare economics writings, in the remainder of this text the term 'utility' will refer to decision utility.

1.2.1.2. Consequentialism in welfare economics

Consequentialism is a key foundation of welfare economics [8]. As a theoretical proposition it stipulates that only utility derived from the outcome of behaviours, such as the consumption of goods or the utilisation of services, is relevant [8,9]. The process by which these outcomes are arrived at is not relevant. To illustrate this point, consider two treatments that return a patient to full health. In Treatment A, due to poor levels of information provision and unhelpful staff attitudes, the patient undergoes an unsatisfactory and uninformed treatment experience, before full health is returned; while during Treatment B no such problems occur. Consequentialism dictates that these two treatments are judged to be equal, based on the assumption that both treatments return the patient to full health, and therefore yield the same utility. Simply put: the end, and not the means, is of importance [27].

The primary critique of consequentialism, and the use solely of *outcome utility*, is contained implicitly within the proposition of *process utility*. Mooney [27] noted that being treated with respect and deference, and maintaining dignity is of concern to patients during the healing process. Furthermore, if a treatment or process reduces the concern or worry a patient experiences then this may be of value, independent of the outcome of treatment. This may be particularly true in areas such as palliative and end-of-life care.

Empirical research, within both general economics and health economics, has sought to determine whether process utility is of importance to the individual. Birch et al. [28] found significant differences in process utility between patients who received aggressive and conservative follow-up treatment after mildly abnormal cervical smears. Benz and Stutzer [29] found workers gained utility from not only the level of pay they received (outcome utility), but also from the way pay is determined through worker/employer consultations

(process utility). Further research by Hahn [30], Frey et al. [31] and Tsuchiya [32] has yielded similar findings. Therefore, the exclusive focus on *outcome utility* may fail to capture large portions of the "effect" a treatment has upon an individual.

1.2.1.3. Individual Sovereignty

The idea of individual sovereignty maintains that the individual is the best judge of their utility or welfare [5,33]. In welfare economics the individual is characterised as an autonomous being, who is rational and has exogenous preferences [34]. It is presumed that the individual has the ability to compare and order alternatives, based on some value of welfare or utility [33]. Through this preference ordering the individual will act to competently maximise their utility through the choices they make [35]. Through this reasoning the individual is maintained as the best judge of the choices required to maximise their utility. For example: when buying a car a consumer will make choices between comfort and style, and speed and reliability. While the salesman can advise, only the individual will know what best satisfies their needs and wants within their budget constraints [5]. Implicitly this rejects the idea of paternalism and the role of a proxy decision maker [35]. It is also of interest to note that the focus of neo-classical welfare economics is not on the individual, but rather on the choices that individuals make. This has led some to criticise the welfare economic perception of the individual as reductionist and restricted, and not adequately representative of the complexity of the human psyche [34].

1.2.1.4. The Pareto principle

To determine whether a social situation is better or worse than alternatives, an assessment must be based on a defined set of criteria, or value judgements [36]. The values and criteria used will determine whether a situation is judged as optimal or not. Welfare economists have

widely used *the Pareto principle*. This says that 'if state A is ranked higher than state B for one person, and all other persons rank A at least as high as B, then A should be ranked higher than B in the social ordering'$_{(pp.2/3)}$[10]. There is judged to be an improvement in social welfare if one person is made better off (moved from state B to A) without anyone else being made worse off (moved from A to B) [4] and under these circumstances change is desirable.

The Pareto principle is considered a weak value judgement [36] for two reasons. Firstly, on the basis that many other value judgements could incorporate it without having it violate their premise and almost any other value judgement of social welfare is likely to violate the basis of the Pareto principle. Ng [36] gives the example of a judgement that makes some people significantly better off, while others are made insignificantly worse off. It is impossible for the Pareto principle, in its strongest form, to incorporate such a judgement. Secondly, neo-classical welfare economists suggest that it is intuitively acceptable to most people. This is a contention disputed by many economists on the basis that when it is emphasised that a change can *only* happen under Pareto conditions, most people would voice concern.

An economic or social situation is said to be Pareto optimal when 'every individual is as well off as he can be made, subject to the condition that no reorganisation permitted shall make any individual worse off'$_{(p.701)}$[22]. Under this condition there is production efficiency (more of one good cannot be produced without producing less of another) and exchange efficiency (commodities cannot be reallocated without making some worse off) [10]. There can be many different distributions of social wealth in which welfare can be considered optimal under this criterion and it is impossible to compare Pareto optimal conditions. Pareto non-comparability also exists in situations where one state is preferred by some, while another state is preferred by others. Therefore, policies which do not offer a uniform/unidirectional increase in social

welfare, but rather redistribute wealth, are Pareto non-comparable. This constitutes a major weakness of this approach.

1.2.1.4.1. Compensation principle

The issue of Pareto non-comparability was addressed by the proposition of *the hypothetical compensation test* [22,37]. This stated that it was not necessary for the economist to demonstrate that nobody suffered from a policy, rather it is sufficient 'to show that even if all those who suffer as a result are fully compensated for their loss, the rest of the community will still be better off than before'$_{(p.550)}$[37]. As its name suggests this compensation doesn't need to be paid in practice, rather it needs to be demonstrated that it could be paid in principle. In doing this Kaldor has separated efficiency issues from equity issues, by delimiting the scope of the economist to efficiency and the responsibility of the politician to distribution and equity [38]. Therefore, it does not, as has been suggested, consider distribution and equity to be irrelevant, but considers it not to be the responsibility of the economist.

This is a stronger value judgement than that contained within a strict interpretation of the Pareto principle. It operates with the understanding that compensation may not be paid in practice and that some may suffer as a result of a policy. It is unclear, when considering sectors such as education and health, how and in what form compensation may hypothetically be paid. Furthermore, it does not allow for the Pareto principle to rank options that are considered Pareto optimal.

1.2.2. Welfarist approach in practice

Effort has been made to "operationalise" welfare economics in health as well as in other fields of research. In health this has taken the form of the cost-benefit analysis.

1.2.2.1. Cost-Benefit Analysis

Cost-benefit analysis offers a theoretically sound form of economic evaluation, underpinned by the theory of welfare economics [39]. Through the identification of all the effects of an intervention or treatment, and the measurement of these effects in a common metric, cost-benefit analysis allows summative benefits to be compared with summative costs [40]. To facilitate this comparison this common metric is usually money [41]. The attachment of monetary values to the cost and benefits of a programme, allows the assessment of whether that programme provides a net benefit to society [39]. In practice, the use of cost-benefit analysis in healthcare has had many practical, conceptual and ethical problems, largely associated with the valuing of benefits in monetary terms [42], in the years since it was first forwarded as an idea [40] and applied in practice [43]. An examination of two approaches to monetary valuation of benefits is the best way to highlight the strengths and weaknesses of this concept. A description of the methodology by which one of these approaches, the willingness to pay approach, elicits values is included

1.2.2.1.1. Human capital approach

The human capital approach uses market wage rates to place monetary weights on gains in healthy time resulting from an intervention or programme [42]. In doing so it assumes that the value of a period of life is equal to the wage that person receives during it. Therefore, the worth of a programme can be valued by the future income that would have been foregone due

to ill health [39]. This approach implicitly assumes that the more a person earns the more their health is worth. This values are then discounted back to yield an accurate present value.

1.2.2.1.3. Willingness to Pay

Willingness-to-pay can be determined by observing individuals or asking individuals for their preferences. In the case of observed preference, an individual's behaviour allows the determination of the value placed on benefits [39]. However, in health care there are a limited number of opportunities to observe this behaviour due to the rare occurrence of illness and the limited presence of markets in the health care system. Therefore, stated preference techniques are often used, whereby people are asked to indicate their preference in monetary terms [39]. Willingness to pay is the primary stated preference technique used in health economics.

The willingness to pay technique elicits the monetary values people attach to healthcare outcomes. The theoretical premise is that the maximum amount an individual states that they are willing to pay (or sacrifice), is an indication of the utility they will gain from that treatment. When considering their willingness to pay, an individual will consider all the attributes of the intervention that is important to them, and not just those designated as important by a third party [42]. Therefore, from the perspective of consumer sovereignty, WTP might be considered a superior measurement process [44]. Theoretically it allows individuals to consider the effects on people other than themselves and, if given enough information about the programme, individuals could potentially consider efficiency and equity aspects. There are a number of methods by which willingness to pay can be estimated. Conjoint analysis and contingent valuation are most frequently used [45].

During the process of contingent valuation participants are initially given information on the hypothetical programme and (possibly) the nature of the illness [46]. They are then required

to indicate their willingness to pay by way of one of three methods. These may either take the form of a response to an open ended question, participation in a form of bidding game (whereby the binary yes/no response to the previous question determines the next value considered), or a hypothetical referendum where the participant is presented with one price for a treatment and they indicate that they either will or will not pay that fee [47].

Conjoint analysis also, normally, requires the presentation of introductory information. Participants are then presented with competing choices described by key attributes, where the levels of the attributes are different in each choice [46]. The process would be repeated numerous times with the levels of the attribute varied. This approach differs from contingent valuation in that it does not require the individual to state their willingness to pay, but rather to choose between two alternatives.

1.2.2.2. Cost benefit summary

Cost benefit analysis theoretically offers a potential method of economic evaluation in healthcare, which would allow for the maintenance of consumer sovereignty and has the scope to measure all aspects of health care important to the individual. However, the use of the sole metric of money means that other forms of evaluation have taken precedence over this technique.

1.2.3. Welfarism critique

Strong critiques have been levelled against the suitability of welfarism as the basis for resource allocation. Welfarism has been critiqued directly as having five primary weaknesses: the incompatibility of the Pareto principle with "real world" situations; the incompetence of the health care consumer; the limiting nature of utility as the basis for evaluation; adaptation; and the monetary valuation of benefit.

1.2.3.1. Pareto principle paralysis

The Paretian welfarist position, that welfare increases when a person, or group of people, are made better with no other person, or group of people, being made worse off, is of little help in allocation decisions [4]. The primary reason for this is that the overwhelming majority of allocation decisions involve winners and losers [7] and the Pareto principle makes no allowance for this. In health care allocation, a treatment that benefits one group will often be funded at the cost of others. Use of the Pareto principle in resource allocation can result in policy paralysis and no decision being taken despite the self-evident need for action. This policy paralysis resulting from the Pareto principle can subsequently leave a situation, which most reasonable people would find sub-optimal, unchanged. 'A state can be Pareto optimal with some people in extreme misery and others rolling in luxury, so long as the miserable cannot be made better off without cutting into the luxury of the rich'$_{(p.32)}$[48]. Furthermore, the compensation principle, is difficult to apply in health, as even theoretically it is difficult to see how "winners" can compensate "losers" with a supplement of "health" or some other utility enhancing attribute [49].

1.2.3.2. The incompetent consumer

The presupposition that the consumer will rationally act to maximise their welfare relies on the assumption of consumer competence [3]. It assumes that the consumer knows all options, understands all options and makes informed and accurate decisions as to which is the best option. However, in health, a lay person is rarely competent enough to decide upon the most efficacious path of treatment. The patient is often not best placed to maximise their utility (or health) and this decision will normally be deferred, to varying degrees, to an agent, in this case a doctor [3,50]. This transferral of competence runs counter to the individualistic nature of welfarism, which cannot consider such expert opinion [5].

1.2.3.3. The utility principle

The consequentialist nature of welfarism holds that only the utility gained from the health resulting from a treatment is relevant. Subsequently welfarism ignores other rich information sources [51] and filters out all non-utility information [52]. Amartya Sen, who is one of the most ardent critics of utility as an evaluative space, questions whether desire fulfilment and being happy is all there is to life [53]. Critics of utility argue that information such as a positive healing experience, ability to undergo treatment that is in line with a person's moral or religious values, or being treated with deference while in care cannot be captured by using utility as the metric of value. Sen has argued that the evaluative space of utility is too thin to form an account of whether a situation offers a social good [54].

In an extension to this line of critique Sen contests that there is a duality in the ethical conception of a person, which utility fails to capture. This is covered in greater detail later in this chapter, but in brief: a person will have well-being goals which they wish to follow and agency goals which they may wish to achieve, with agency referring to goals, commitments

or values that can, but do not necessarily, include well-being. Agency goals can both conflict and complement one's well-being. Under Sen's interpretation of utility, welfarism is not able to capture both well-being and agency, and with this inability, information of real importance is lost.

In his final strand of critique, Sen somewhat switches the focus to the question "equality of what?" [55]. Sen notes that every enduring social theory has demanded equality in some space [53]. Over time, theories and social arrangements have proposed equality in the form of liberties [56] and primary goods [57], as well as in many other areas. Utilitarianism seeks equality in the evaluative space of utility: everyone's utility gains hold the same weight, and a utility loss is acceptable to no one. But, demanding equality in one space, can lead to or justify inequality in another [55]. Therefore Sen argues that the question "equality of what?" is of utmost importance [53]. Sen rejects utility as the appropriate evaluative space on the basis that seeking equality of utility, especially in conjunction with the Pareto principle, can cause unacceptable inequality in another space [53].

1.2.3.4. Adaptation and outcome assessment

The potential for people to adapt to their situation or surrounding is considered a further limitation to using utility as an outcome metric [53]. In everyday life, adaptation is a part of the human psyche that prevents individuals from becoming transfixed on details and gives them the ability to react appropriately to future events. However, in outcome assessment it may result in inappropriate measurement, which can be compounded by using utility as an outcome metric.

> *'A person, who has had a life of misfortune, with very little opportunities, and rather little hope, may be more easily reconciled to deprivations than others reared in more fortunate and affluent circumstances. The metric of happiness may therefore distort the extent of deprivation, in a specific and biased way'* $_{(p.45)}$ *[58]*

Sen [55] argues convincingly that in the case of poverty the metric of individual desire fulfilment, used by welfare economists, may not adequately measure a person's deprivation. His assertion that there are those who have accepted the hardship that exists in their lives, and tapered their expectations so, is illustrated in his editorial for the BMJ [59]. Sen observed higher reported morbidity in an affluent Indian state, which had higher life expectancy, than in the deprived state of Bihar, with lower life expectancy. Sen cites, among other things the people of Bihar's 'very low perception of illness'$_{(p.861)}$ as a reason for this disparity in reported morbidity.

Work in health and disability has shown similar differences. Seminal works by Sackett and Torrance [60] and Brickman et al. [61] showed, respectively, that there were considerable differences in how dialysis patients and the general public valued health-related quality of life for those requiring dialysis and that paraplegics were only marginally less happy than non-paraplegic controls. Oswald [62] has recently presented longitudinal evidence that the degree of adaptation of 'life satisfaction' is in the order of 30% to 50% in those who suffered "quite serious" levels of impairment.

1.2.3.5. Monetary valuation

The application of the welfarist approach to resource allocation through the cost-benefit analysis model has a number of weaknesses and limitations. Firstly, the human capital approach raises ethical questions: should the worth of an individual's health be valued solely by their financial activity? Does an unemployed person's health have no value? Furthermore, practical limitations are also present. Wage rates only accurately represent productivity when certain labour market conditions are met [39] and any imperfections within the labour market, of which there are normally many, will give inaccurate quantifications of productivity. If the

lost employee can be quickly replaced from a pool of suitably skilled people, the productivity loss may be limited to a "frictional cost" of recruitment and training [42]. Lastly, and possibly most damning, is a criticism which comes from within welfare economic theory: this approach is not based on individuals' preferences or valuation of health gains.

Willingness to pay is inevitably a function of ability to pay$_{(p.211)}$[63], leading to equity concerns [42]. A further concern surrounds whether asking people to value health gains in monetary terms could invite methodological problems. Individuals may find it unacceptable, or incomprehensible, to be asked to decide how much they would be willing to pay for health care, or how much they are willing to accept for continued poor health or death. This may lead to protest bids, whereby the respondent gives a value that indicates an objection to the question, i.e. an impossibly high figure, or a zero bid [42].

1.2.4. Welfarism critique summary

The critical appraisal of the welfarist approach has led many to believe that it is poorly fitted to resource allocation, especially in health [4,64]. This critique has been on both practical and theoretical grounds. While some have maintained the ability of welfarism to assist in resource allocation [65], the majority accept that a different approach is needed [4,7,48,66]. This led to the development of extra-welfarism.

1.3. The Extra-Welfarist Approach

This section describes the main motivation for the development of the extra-welfarist approach. The primary differences between extra-welfarism and welfarism are described and the ways in which it has been put into practice are examined. The section concludes with a summary of the critiques levelled at the extra-welfarist approach, which often focus on its practical application.

1.3.1. The seeds of extra-welfarism

Brouwer et al. [9] have identified a number of 'seeds' that led to the rejection of welfarism and the acceptance of the need for a different model of resource allocation in health. Brouwer's "seeds" are non-direct critiques of welfarism.

The first seed is the concept of the "merit good". Some goods are too meritorious to be left open to the will of market mechanisms, and should be subsidised, in some form, by the state [5,67]. The concept of a 'merit good' can be understood as a good which has a greater value to society than is reflected in its market price, due often to the externalities that the good produces [67]. If provision was left to market forces they would be under provided, so normally they are subsidised by the state. Education and health care are routinely considered as merit goods, as is transport infrastructure such as roads.

A second seed was the assertion of a specific, rather than general, desire for equality. Tobin [68] proposed that people consider that the allocation of some goods and services should be based on egalitarian principles, while for other goods this desire is absent. For example; people may accept a situation where some people do and some people don't have access to Perrier, but not a situation where some don't have access to basic, clean drinking water.

Similar suppositions are found in John Rawls's [57] 'basic goods' theory and Walzer's [69] spheres of justice premise.

Thirdly, great weight was added through Sen's offer of functionings and capabilities as more appropriate metrics for the measurement of wellbeing [30,48,53,70,71]. The capability approach has found traction as an idea in policy areas outside of poverty and human development, where it was initially proposed. A fuller description of this approach, which contains a strong critique of welfarism, is provided in a later section of this chapter.

The fourth seed, and possibly the most damning criticism of the welfarist approach from a practical health provision perspective, is the continued rejection of welfarist policies in health by governments. In the UK, the 1944 white paper titled 'A National Health Service' and in the US, the Social Security Act of 1965, showed the rejection of strict welfarist policies [9]. The majority of governments are focused on ensuring a basic level of healthcare for all citizens. An understanding of trades offs, where some lose and some win, is essential for ensuring this provision. Furthermore, provision of health is not left solely to the individual and there is an acceptance that decision makers can be a good addition to health resource allocation [72].

Extra-welfarism was developed throughout the 1970's and 1980's in response to critiques similar to the one presented above [5,72,73]. Coast et al. [4] detail extra-welfarism's large theoretical departure from standard neo-classical welfarism:

> *'extra-welfarism is defined as transcending traditional welfare by supplementing these welfares with other 'non-goods characteristics' of individuals such as health state, freedom of choice and even quality of relationships between individuals'* *(p.1192)*.

1.3.2. Theoretical basis of extra-welfarism

The meaning of extra-welfarism has caused considerable confusion and consternation since its beginnings [33,74]. Culyer's extra-welfarism [73] draws heavily on Sen's [70,75,76] theory of functioning and capability to enrich the evaluative space, by permitting outcomes other than utility and allowing these to be elicited from people other than affected individuals [5,9]. It has been both defined as surpassing traditional welfare economics [73] and criticised for adding little content to already existing theories [77]. This confusion has been so chronic and persistent that as recently as 2008, Brouwer, Culyer and colleagues [7] have further tried to clearly delimit extra-welfarism, through a clear statement of ways in which it diverges from the welfarist approach. Brouwer et al. [9] assert that the extra welfarist approach differs from welfarism in four ways. Firstly, it 'permits the use of outcomes other than utility'$_{(p.330)}$[7] in the analysis of the welfare of the individual. Secondly, it does not constrain the sources of valuation solely to affected individuals. Thirdly, it allows outcomes to be weighted by principles other than preferences. Finally, it departs from the Pareto principle in allowing interpersonal comparisons in the evaluative space.

1.3.2.1. Beyond utility

The inability of the welfarist approach to admit non-utility information has been a focal point for criticisms of welfarism. Sen [78] proposed information on the basic capabilities of an individual as an example of non-utility information that is important and relevant to the assessment of a person's welfare and asserts that people have needs that are independent of their utility.

Extra-welfarism relaxes the assumptions of welfare economics [77] by rejecting the sole focus on utility-based interpretations of welfare [79]. In doing this it allows information other than

utility to be included in the assessment of welfare. This information can include the process of care and, importantly, health. In extra-welfarism health is not valued for the utility it produces, rather for its own sake, on the basis that health is the principal outcome of health care [64]. In admitting other sources of information, the extra-welfarist is not rejecting utility information, rather recognising that utility is one of many sources of information [7]. Nor are the preferences of the public necessarily superseded. Preference measurement can be applied to non-utility information, such as health gains or capability improvements [80]. This movement beyond utility has allowed the claim that, as a theoretical model, extra-welfarism offers more breadth than neo-classical economics [7,75,80].

1.3.2.2. Accepting paternalism

Extra-welfarism departs, in two ways, from viewing affected individuals as the only relevant source of evaluation. Firstly, it recognises the role of a decision maker in defining the evaluative space [7]. This requires the selection of what should be, and what should not be, considered relevant by a policy or decision maker. Sen [81] claims that this is not an embarrassment for extra-welfarism; a claim which is reaffirmed when we consider that Sugden and Williams had previously proposed such an approach, under the welfarism umbrella [72]. Secondly, extra-welfarism permits "stakeholders", other than the affected individuals to be regarded as appropriate sources of information. Therefore family members, doctors, health managers and citizens' juries can all be considered legitimate sources of information.

1.3.2.3. Weighting of outcomes

Support has been found for the distribution of health outcomes by some method other than sum maximisation [82], such as distributed in favour of the young [83], in favour of lower

socio-demographic levels [84,85] or in favour of those who have the lowest levels of existing health [86]. While there is no consensus among extra-welfarists on if, and on what terms, health should be distributed, the extra-welfarist model would theoretically allow for such prioritisation. Outcome measures could be used to inform such judgements. Outcome gains for those below a defined level could be given greater weights than those above that level. Culyer [74] notes that it is possible to fully integrate these weights into cost-effectiveness analysis, allowing both efficiency and equity to inform health policy and resource allocations.

1.3.2.4. Interpersonal comparisons

Interpersonal comparisons are the corner stone of the extra-welfarist approach to decision making [7]. The use of health measures allows individuals to be compared on some measure of health. This can facilitate comparison between patients with different conditions and funding decisions between two or more very different treatments. Therefore, the policy paralysis, which occurs from the inability of the Pareto Criterion to incorporate interpersonal comparisons, is avoided.

1.3.3. Extra-welfarism in practice

1.3.3.1. Cost-effectiveness analysis

Cost-effectiveness analysis offers a method by which both the costs and the consequences of treatments can be measured within the theoretical boundaries laid down by the extra-welfarist approach [87]. Cost-effectiveness analysis quantifies the benefit of a treatment through a natural single unit, such as lives saved, cases detected or life years gained [88]. A number of outcome measures or endpoints are suitable for use, with the outcome expressed in cost per unit of effect [87]. Under cost-effectiveness analysis, intermediate endpoints, such as blood pressure as an indicator of stroke risk, and final endpoints, such as cardiovascular related

death, or all-cause mortality, can be used. Although intermediate outcomes may be acceptable, ideally a link between intermediate and final outcome should be clearly demonstrated [87]. Patient completed quality of life measures are a frequently used outcome measure in cost-effectiveness analysis. A summary of these measures is provided in chapter two.

1.3.3.2. Cost-utility analysis

Cost utility analysis is the most frequently used form of economic analysis for decisions involving health care resource allocation [89]. Here utility refers to the preferences of groups of individuals for a health outcome. This method allows health outcomes to be 'valued according to their desirability'$_{(p139)}$[89]. Therefore, while cost-utility analysis is often considered as a subgroup of cost-effectiveness analysis, it has clear and important distinctions, not least that it recognises the importance of considering public preferences [90].

The methods by which these preferences are elicited vary, but focus around four main processes for drawing out preferences: the visual analogue scale, standard gamble, time trade-off and the discrete choice experiment. The visual analogue scale is a simple approach whereby subjects are asked to rank their preferences in order of most to least preferred and then to place the outcomes on a scale so that the distances between placements correspond to the difference in preferences [89]. If options were only marginally preferred then spacing would be small, whereas clear preferences would see larger spacing. In the standard gamble approach the subject is offered two options. In option one the subject stays in a chronic state for life; option two there are two possible outcomes, with probabilities attached: outcome one the person is returned to full health and lives for as set number of years (probability *y*) and option two the subject dies immediately (1-probability *y*) [89]. The probabilities are varied,

systematically, until the participants are indifferent between the two options. Third, in the time trade-off approach the subject is offered two options [89,91]. Option one is to live in the diseased state for a given period of time. Option two is to live in a healthy state for a period of time less than Option one. The time is varied until the subject is indifferent. Finally, a discrete choice experiment is a survey method based on choices of attributes [92,93]. The participant is presented a number of different hypothetical states. In each scenario the participant is asked to choice between two or more options which vary on important attributes or characteristics [92–94].

These processes through which preferences are elicited are time consuming and too cumbersome for use in the majority of research settings, where efficiency and speed of data collection is required. In response to this limitation, pre-scored, patient-completed, health-related quality of life measures are routinely used. In this situation, weightings for the responses to a health-related quality of life measure are elicited through one of the methods described above and "attached" to the measure. The response to the quality of life measure can then be scored, using the weighted responses, to give an output that recognises patient or the general public's preferences. These quality of life measures are considered in greater detail in chapter 2.

1.3.3.3. The quality adjusted life year

The quality-adjusted life-year (QALY) was introduced in 1968 [95] and furthered by Weinstein et al. [96] and others [97]. The QALY defines health as value weighted time accumulated over a given time horizon [98]. The QALY represents a common unit of effectiveness that can be used across economic evaluation in different clinical areas [90]. Using the QALY, a year of life is adjusted for the life quality during that year and reduced

down to a single value [97]. A score of 1 indicates that someone is living in perfect health for one year. A score lower than 1 suggests that a person is either living in a degree of poor health for that year or they have lived for less than a year. A treatment or programme can increase QALYs by either increasing the quality of life during a period of time, or extending the period of time a person lives for. The QALY concept dictates that quality weighting is preference-based, anchored at death and measured on an interval scale [89]. The QALY incorporates both the quantity of life (mortality) and the quality of life (morbidity) into one measure. The QALY seeks to incorporate preferences through use of the cost-utility approach to obtain the weighting of the measurement.

1.3.4. Extra-welfarism critique

The critiques levelled against extra-welfarism have been focused on the practical application of the approach rather than the theoretical proposition. In practice, extra-welfarism's complexity and breadth is not retained and a more limited departure from welfarism is realised [5]. Three main limitations are noted: the limitation of the evaluative space, the limitation of the evaluative scope and the commitment to maximisation.

1.3.4.1. The limitation of the evaluative space

Theoretically, extra-welfarism expands the evaluative space by allowing outcomes other than utility to be used in evaluation [7]. In practice the expansion of the evaluative space in extra-welfarism is more limited, with utility replaced with health as the sole outcome measure [77]. Therefore, rather than permitting other outcomes, extra-welfarism has changed the exclusive focus from utility to health. Furthermore, while the health of patients is viewed by many as highly relevant to the evaluation of health technologies and treatments, it could be perceived as a narrower evaluative space than that of utility [4,49].

The adoption by extra-welfarists of health as the sole outcome measure, may be considered especially limiting in some life situations and for some fields of medicine [99]. For those who are in the end stage of life or living with a chronic disease, health may not be the only, or even the most important, outcome of their treatment. Furthermore, for medical specialties such as geriatrics, or in situations where social and medical considerations overlap, such as in social care or public health, use of health as the only measure of worth for the treatment being provided is unlikely to capture the effect of a treatment accurately .

1.3.4.2. The limitation of evaluative scope

Along with the limitation of the evaluative *space*, a less discussed limitation is in the evaluative *scope* of the approach: the people considered relevant. Cost-effectiveness analysis has maintained a focus on patients, on the assumption that the patient population will receive any health gains provided. This assumption excludes the potential that close relatives and informal carers may benefit from health and non-health gains in the quality of life as the health of the person they care for improves [100]. Currently very few economic evaluations go beyond the patient; however discussion is taking place about the appropriate scope of extra-welfarist analysis [101,102]. Furthermore, evidence exists that this limitation of the scope may lead decision makers to reach different decisions on the cost-effectiveness of an intervention than would be reached if the broader impacts of an intervention were considered [102].

1.3.4.3. The commitment to maximisation

The practical expression of extra-welfarism has maintained the importance of maximisation as a measure of good, with the maximisation of health replacing utility [103]. Coast [4] states that this is not linked to any theoretical basis, rather it has been asserted by early decision makers and reported in early academic works [96,103]. Potentially, this occurred as it is a position which is acceptable or traditional within economics and can be viewed as a weak value judgement.

This commitment to maximisation has been further criticised for being against the wishes of the majority of the public, who would like to see a more egalitarian distribution of health benefit and gain. Dolan et al's [82] review of the literature predating 2001, showed under a number of circumstances, that the public favoured a distribution of wealth by some criteria

other than maximisation of population health. A large body of literature has been published since 2001 which has showed that the public may want to prioritise the young over the old [104,104] and the severely sick over those who are less sick [86,105], while there is some indication that the public wants to consider whether the illness was self-inflicted in the decision making process [106]. Taken as a whole, the literature suggests that under some circumstance the public do not support maximisation of health.

1.4. The capability approach

This section will describe the theoretical underpinnings of the capability approach. Many of the previous sections of this chapter (notably sections: 1.2.3. Welfarism critique, 1.3.1. The seeds of extra-welfarism, 1.3.2. Theoretical basis of extra-welfarism and 1.3.4. Extra-welfarism critique) are important in understanding the capability approach and should be considered in conjunction with the following section. The section concludes with a summary of the challenges of "operationalising" – putting into a health-research-useable form – the capability approach. A full summary of current efforts to operationalise the approach and an in-depth description of the ICECAP-O and ICECAP-A measures are provided in Chapter 2.

1.4.1. Theoretical basis of Capability

'The capability approach is a broad, normative framework for the evaluation and assessment of individual well-being'$_{(p.94)}$[1]. Proposed by Amartya Sen and developed by Sen, Nussbaum [107] and others, the approach offers a substantial diversion from welfairsm and the practical expression of extra-welfarist approaches. Generous credit is given by Sen to theorists, economists and philosophers, ranging from Aristotle to Adam Smith to John Rawls, who preceded him and influenced and shaped his thinking [55]. The capability approach proposes a rich and broad evaluative space on which judgements should be based [53,54]. The capability approach has been considered for use in health research and health service evaluation because of both the limitations with the current methods of assessment and the increasing recognition that health interventions often result in outcomes other than health.

1.4.1.1. Functioning and capability

The central feature of Sen's capability approach is to advocate the evaluation of situations based on the extent to which a person is able to function [66]. Sen [53,75] states that the

ability to achieve valuable functionings, consisting of "beings" and "doings", determines a person's well-being. Sen defines actual achievement as "functioning" (e.g. living without illness). A person's achieved state of well-being can be assessed through a functioning vector, which is a measure of a person's functionings [53,54]. Sen defines freedom to achieve as "capability" (e.g. the option to live without illness). When assessing well-being Sen advocates going further than measuring a person's achieved functionings. He says that measuring the functionings which a person has the ability to achieve (their capability) offers a fuller assessment of quality of life [53,70,75]. A person's capability set can be thought of as a set of vectors of functionings from which a person can choose.

The benefits of measuring capability rather than functioning can be described using a slightly altered version of Sen's well-known fasting example [70]. Consider two people, Person A and Person B, who both require an operation and consequently a blood transfusion. Person A is not able to have the operation because there is no blood available in her country or geographical area. Person B is not able to have the operation because of an objection (religious or otherwise) to receiving blood, despite plenty of blood being available in his country. In assessing these individual's functionings one would conclude that they are the same: neither had the operation. However, by assessing these individual's capability they are shown to be very different: Person B, in an objective sense is able to have the operation, whereas Person A does not have that choice. Sen states that there is 'at least no informational loss in seeing well-being assessment in terms of capabilities, rather than directly in terms of achieved...functioning'$_{(p.39)}$ [70].

1.4.1.2. The relationship between a good and an individual

A motivation for measuring well-being in terms of functionings and capabilities is Sen's assessment of the relationship between financial wealth, physical goods and an individual. Being well off is not one and the same as being well [52]. Well-being is influenced by a person's income, but a person's income does not determine a person's well-being. Personal characteristics ensure that while there is a causal relationship between money and well-being, this relationship is not complete. For example, a person with severe mental or physical disability will require greater wealth and will need to possess more goods to be able to live in the same quality of life as a non-disabled individual [52].

Sen makes the distinction between four different aspects of the relationship between a good and an individual, by using the example of a bicycle [108,109]. The good is the bike; the characteristics are the qualities of the bike (e.g. movement); the utility is the pleasure gained from owning the bike; and the functioning is an individual's use of the bike (riding) [108]. The personal characteristics of an individual will determine both the utility gained from the bike and the functionings achieved: a competent and enthusiastic cyclist would gain more from the good than an individual who cannot cycle.

The variations in personal characteristics lead to variations in the conversion of good, resources and wealth into both utility and functionings. Therefore, resources do not have an intrinsic value, rather their value is in the opportunities they provide [54]. The capability approach sets out to value these opportunities directly, and in doing so accounts for the significant effects that the conversion of goods and resources can have.

1.4.1.3. The distinction between agency and well-being

As noted earlier, the capability approach advocates a richer and broader evaluative space than traditional welfarist and extra welfarist approaches. The capability approach identifies two distinctions, which form four subcategories within the evaluative space. The first, discussed above, is between capability and functioning. The second distinction is between "well-being goals" and "agency goals" [70].

Well-being goals are basic goals that relate directly to the well-being of that person. These may include living without severe illness, living under shelter and having acceptable nutrition [70,110]. Agency goals are normally more complex goals that an individual has reason or motivation to seek [70]. These goals can be ambitions such as to live by a certain set of standards or morals, or career ambitions. Agency goals may be linked to, or may themselves *be* well-being goals [53]: a person may enjoy keeping themselves healthy and free of disease by eating well and taking regular exercise. The pleasure gained from this is greater than simply the basic achievement of a minimum standard of life: it has greater agency to the individual. However, agency goals may also be goals that are different to, and have a deleterious effect upon, well-being [53]. For example, a person may be highly career focused and work long, stressful days in order to achieve the promotion or position they value. This work may take its toll upon their health, and therefore they are achieving their agency goal at some cost to their well-being.

These two distinctions, between capability and functioning and between well-being and agency, result in four sub-categories in the evaluative space: well-being achievement (functioning); agency achievement (functioning), well-being freedom (capability) and agency freedom (capability) [70].

1.4.1.4. Basic versus complex capabilities

Sen [70] makes a further, two level distinction between basic capability, 'the ability to satisfy certain crucially important functionings up to certain minimally adequate levels'$_{(p.41)}$, and more complex capabilities, which is closely related to well-being and agency. Sen [53,76] states poverty can be viewed as a failure of basic capabilities, such as the ability to be disease free and have basic nutrition, to reach acceptable levels. In such circumstances, a focus on basic capabilities and functionings may allow us 'to go a fairly long distance'$_{(p.44)}$[53] in terms of analysis. However, in situations where these basic capabilities are taken for granted, such as in many developed countries, then a broader range of more socially orientated capabilities may be needed to judge a person's well-being and agency [76]. In the analysis of health and social care interventions in a developed country, a broader range of capabilities would likely be required.

1.4.2. Operationalising the capability approach

The capability approach has a number of notable differences from other, competing, normative theories. The approach is not a list of commodities or personal traits and Sen has purposively left it incomplete to allow for plurality in the evaluative space [110]. This makes capability a complex theory. Sugden [54] has questioned the extent to which the theory is operational: 'given the rich array of functionings that Sen takes to be relevant, given the extent of disagreement among reasonable people about the nature of a good life, and given the unresolved problem of how to value sets, it is natural to ask how far Sen's framework is operational'$_{(p.1953)}$. In some ways Sen has compounded this challenge by stating that 'if an underlying idea has an essential ambiguity, a precise formulation of that idea must try to capture that ambiguity rather than hide or eliminate it'$_{(pp.33-34)}$ [70].

A number of proposals have been given regarding how to operationalise the approach. Robeyns [111] describes a continuum of ways in which the capability theory could be made operational. At one end of the continuum is structured, methodological driven research where a reductive statistical technique is used upon a rich data set to identify important functionings and capabilities. At the other end is an approach based on theoretical underpinnings.

Nussbaum's list of "central human capabilities" is the most well-known example at the theoretical end of the spectrum. Nussbaum proposes ten capabilities (life; bodily health; bodily integrity; senses; emotions; practical reason; affiliation; other species; play; and control) which she holds all humans to be morally entitled to [107]. Critiques of this list have stated that the list lacks legitimacy and consensus. Nussbuam states that the list is formed at the abstract level and should be translated into implementation at the local level [111], but many of the ten capabilities on the list might be considered irrelevant to many research settings, or it might be considered unreasonable to expect a medical intervention to alter some of the capabilities listed.

Sen has not proposed a list of capabilities and has indicated that in different contexts different capabilities are likely to be important [4]. In line with this thinking Robeyns proposed a "check and balance" procedure by which capabilities should be selected by researchers and policy makers. This procedure is: explicit formulations, where the list is described, discussed and defended; methodological justification, whereby the method of formulation is discussed and defended; different levels of generalisability, whereby the lists are generated at stages ranging from ideal theory to practical implementation; and exhaustion, where checks are completed to ensure no important capability is left out.

Despite the challenges, the capability approach has been recognised as providing a theoretical framework through which disability can be described [112] and assessed [110]. It has been cited as a way through which the effects of public health interventions, which may reach outside of a purely health domain, can be accurately assessed for health economic evaluation [113]. Furthermore, the potential for the approach to provide a richer evaluative space for economic evaluation in health per se and to challenge some of the value judgements which currently exist in health economic evaluation has been noted [4,66].

1.5. Conclusion

In this chapter the central tenets of welfarism, extra-welfarism and the capability approach are introduced. The theoretical evolution from welfarism towards and extra-welfarist approach has been described and the limitations of both approaches are discussed. Welfarism critiques have remained largely theoretical and have focused on the potential results of using Utility as the evaluative space in conjunction with the Pareto Principle; while for extra-welfarism critiques have been centred on the truncation of the scope of the approach when put into practice in health.

The capability approach may provide a solution to many of the limitations of the welfarist and extra-welfarist approaches. It offers a broader evaluative space than health or utility, which may provide a fuller conceptualisation of the value of a given situation. This approach is the focus of this thesis, which seeks to extend recent work that has developed two new patient reported outcome measures for use within a health and social care setting: the ICECAP-A [114–116] and ICECAP-O [109,117,118] measures.

CHAPTER 2. QUALITY OF LIFE AND CAPABILITY MEASUREMENT IN RANDOMISED CONTROLLED TRIALS

2.1. Chapter introduction

Chapter 1 provided a description of how, theoretically, the capability approach diverged from welfarist and extra-welfarist approaches. The practical application of the capability approach will likely differ from established methods for measuring quality of life in health research. Attempts to put this theory into practice in a health research setting are at an early stage. In order to provide an indication of how the capability approach may extend current methods for assessing quality of life, a summary of the types of quality of life measures used in health research, the purposes of use and how measures "fit" into extra-welfarist theory is provided in this chapter. A description of efforts to operationalize the capability approach in a research setting is provided with reference to the OCAP and OxCAP measures, the ASCOT measures and the ICECAP capability measures. The development and valuation of the ICECAP measures are discussed in detail. A methodological review of studies assessing the validity of the ICECAP measures concludes the chapter. First, however, this chapter starts with a summary of a modern trial. Randomised controlled trials are considered a gold standard research technique [119]. Health economic analysis is frequently conducted alongside trials [120] and quality of life measures are increasingly being utilised as outcomes in trials [121,122] [119]. If the ICECAP measures are to be used routinely in health research, they will be used in trials. Two randomised controlled trials are also the vehicle through which this thesis has assessed the validity of the ICECAP measures. A summary is provided through discussion of the characteristics of the modern trial methodology: randomisation, blinding and comparison.

2.2. The randomised controlled trial

A central feature of modern medicine is the use of the randomised controlled trial in evaluating the effectiveness of medical procedures and interventions [123]. A rigorously planned and properly executed 'prospective study comparing the effect and value of intervention(s) against a control'$_{(p.2)}$ is a powerful investigative tool, which is considered a gold standard experimental technique [119]. Modern trial methodology is a complex discipline that is focused around three basic characteristics: randomisation, blinding and comparison [123].

2.2.1. Characteristics of a randomised controlled trial

2.2.1.1. Randomisation

Randomisation, the process by which a patient entering a trial has a pre-defined chance of being allocated to the treatment or control group, is a key feature of the modern randomised controlled trial [124]. It has three main methodological advantages. The deliberate introduction of an element of chance into the assignment of treatments reduces the possibility for bias in the allocation of treatments [125,126]. It helps produce groups that are similar both by factors that will be measured by the trial, such as health status, age and ethnicity, and factors that are unobserved, such as undiagnosed disease. The production of groups that are similar means that they are also statistically comparable [126]. Finally, it safeguards the validity of statistical tests [124]. For example, use of the chi-squared and t-test can be justified on the basis of randomisation alone. If, however, the sample is not random a number of other considerations, such as the distribution of responses need to be considered [126].

The process of randomisation can be very simple, such as the toss of a coin or the use of a random number generator[127]. While such unrestricted randomisation is an acceptable approach, other methods of randomisation do have practical advantages for a trial. Block randomisation, where participants are randomised in groups, can reduce the risk of serious imbalance due to chance [125,128–130]. Stratified randomisation, whereby participant characteristics are considered in order to ensure groups that are comparable by some prognostic or risk factor, can be useful in small studies where a potential imbalance is likely [127]. Whatever the process of randomisation chosen, the randomisation schedule should be concealed from both the researchers and participants of the trial, to ensure the study are "blinded".

2.2.1.2. Blinding

The purpose of blinding, or the concealment of the treatment allocation, is to 'prevent the identification of the treatments until all such opportunities for bias have passed'$_{(p.1914)}$[125]. There are a number of forms of bias that can occur when the treatment allocation is not concealed [131]: response bias is when participants respond according to how they think they are expected to (e.g. intervention group should improve); attrition bias is the increased dropout amongst patients who know they are not receiving the intervention treatment; and outcome or observer bias is the potential of assessors to preferentially assess patients who they know are receiving the intervention. Furthermore, blinding removes the potential for a placebo effect or the psychological impact of patients knowing that they are or are not receiving the intervention.

A double blind trial is a study where neither the participant nor the researchers (both those collecting and analysing data) know the treatment allocation [125]. A double blind study

design is considered the “gold standard” in trials. While methodologically superior, double blind studies can be very difficult in practice, especially in trials where researchers are delivering different treatments or interventions. A single blind study is a study where only the investigator knows the allocation and an unblinded, or open label, study is a situation where both the researcher and the participant know the allocation [119]. The risk for bias in this last situation is high; however, in some situations, such as surgical trials, this is the only practical option [132].

2.2.1.3. Comparison

The comparator treatment is an important methodological and ethical consideration. If an established “best” therapy exists the control group in a trial would usually be expected to have this treatment administered. In situations where the best therapy is not widely available, either for cost reasons, lack of clinical ability (often in the case of surgical interventions) or organisational capacity, then placebo controls may be considered. Use of a placebo control when a best therapy exists is considered ethically questionable [119].

A trial which has the objective to show that the new intervention or therapy is more effective than the control is termed a superiority trial [125,133]. However, not all trials are superiority trials. Non-inferiority trials have the objective of a comparison with the control, to see if the intervention treatment is not clinically inferior to the control [125,134]. This comparison may be needed if the new treatment is thought to be cheaper than the existing treatment, or to induce fewer side-effects.

2.3. Measurement of outcomes for cost-effectiveness analysis alongside clinical trials

Cost-effectiveness analyses express value in cost per unit of effect, with the measure of effect being chosen to suit the purpose of the intervention under consideration [88]. For example, a screening programme may assess cost per disease detected, while a curative intervention may assess cost per disease cured. Early cost-effectiveness assessments used a number of different measures of effect [89]: reduction in mmHg of blood pressure in a cost-effectiveness study of treatment of hypertension [135]; cases of deep-vein thrombosis detected in a cost-effectiveness analysis of a screening programme [136]; or episode-free days in the clinical area of asthma [137].

Use of such clinical or "natural" outcome measures have some limitations. They make comparison between cost-effectiveness interventions in different areas very difficult. Furthermore, trade-offs cannot be made explicitly [89]. Consequently quality of life instruments have become an increasingly important way of assessing the efficacy of an intervention or treatment in a randomised controlled trial [119]. They are used both as outcomes in health economic analysis alongside the trial and as primary and secondary measures independent of health economic analysis.

The development of quality of life measures can be traced back to attempts to extend assessment beyond clinical outcomes, such as the presence or absence of disease or disease markers [138]. One of the earliest efforts, the Karnofsky Performance Scale, was conceived in 1947 as a one item measure, which asks patients to rate their health on a 0 to 100 scale [139]. This was later followed by more detailed, multi-item assessments of health status, such as the Sickness Impact Profile [140] and the Nottingham Health Profile [141]. Later

measures, such as the EuroQoL five dimension (EQ-5D) [142,143], Short Form (36) Health Survey (SF-36) [144] and European Organisation for Research and Treatment of Cancer quality of life questionnaire (EORTC QLQ-C30) [145], while maintaining a focus on health, have included assessment of social, psychological or emotional functioning.

2.3.1. Types of health-related quality of life measures

The health-related quality of life measures currently used in health research can be categorised as: generic, illness specific, aspect specific and patient specific.

2.3.1.1. Generic health-related quality of life measures

Generic measures can be used in studies of different illnesses and differing patients. Many of these measures may be applicable for use within the general population. Two frequently used examples are the EQ-5D and SF-36.

The EQ-5D is a brief measure which assesses the quality of life attributes of mobility, self-care, usual activities, pain/discomfort and anxiety/depression [142,143]. Two versions now exist with the earlier version offering three response categories (EQ-5D-3L) and the later five (EQ-5D-5L) [146–148]. From these questions a single index score can be calculated.

The SF-36 is a longer questionnaire than the EQ-5D-3L and addresses eight health attributes: physical functioning, physical role, bodily pain, general health, vitality, social functioning, emotional role and mental health [144]. A score for each attribute is provided by a scale. From these scales, two summary scores of physical health and mental health can be calculated.

2.3.1.2. Illness specific health-related quality of life measures

An example of an illness specific questionnaire is the EORTC QLQ-C30 cancer specific questionnaire [145]. The measure was designed for use across a range of cancers. This 30 item questionnaire assesses physical, cognitive, emotional and social functioning, broad symptoms, such as fatigue and pain, and specific symptoms, such as dyspnoea and diarrhoea. In addition to the core QLQ-C30, the EORTC group also offer a number of additional modules that recognise that different cancers have different morbidities and different treatments have different side-effects [149].

2.3.1.3. Aspect specific health-related quality of life measures

Aspect specific questionnaires are used to assess the effect of a disease or treatment on an aspect of quality of life that a researcher wants to explore in more detail [138]. An example of such a measure is the Hospital Anxiety and Depression Scale (HADS). The HADS is a 14 question measure that has been used across a wide variety of clinical conditions [150]. It assesses both anxiety and depression, and produces subscales for each. Scores over a cut off are indicative of anxiety disorder or depression.

2.3.1.4. Patient specific health-related quality of life measures

A small number of measures have been developed based on the characteristics of the patients. For example the PedsQL is a short generic, standardised instrument which is appropriate for use with children and adolescents and their parents [151]. The Geriatric Depression Scale is an aspect specific measure appropriate for use with elderly patients as a measure of their psychological health [152].

2.3.2. Health-related quality of life measures and extra-welfarism

The majority of quality of life measures used in health and clinical research still focus on health functioning [138]. In a number of cases, measures that were designed as health status measures are frequently referred to as quality of life measures in the literature and an assumption is often made that if a person's health is poor, then quality of life must be reduced [138]. Health is used as an indicator of quality of life. Therefore, the majority of instruments used can, and are, most accurately described as health-related quality of life measures.

Cost-effectiveness analysis seeks to quantify the benefit of a treatment through a single unit of effect [88]. This unit may be life years, or cardiac events. It may also be scores from patient-reported health-related quality of life measures. The outcome measures discussed above are a selection of measures which "fit" within the theoretical boundaries of cost-effectiveness analysis: they offer quantification of benefit through the score which they provide. The measures also "fit" within the practical application of extra-welfarism seen in the literature, which has sought to focus on health and health functioning.

Cost-utility analysis differs from cost-effectiveness analysis in requiring health outcomes to be valued according to the utility they provide. This measurement of value, or preferences, allows the formation of the QALY, or some variant, such as the DALY [89]. The three most widely used techniques, which were discussed in section 1.3.3.2. Cost-utility analysis (Time Trade Off, Standard Gamble and Visual Analogue Scale (VAS)), are all time-consuming methods of eliciting preferences. Pre-scored measures have become a popular method for "bypassing" these time-consuming measurements each time an economic analysis is conducted [89].

2.3.2.1. Utility measures

Three commonly used health-related quality of life measures with preference scores attached are the EQ-5D-3L [142,143], the Short Form 6 Dimension (SF-6D) [153,154] and the Health Utilities Index (HUI) [155,156]. The EQ-5D-3L, the descriptive system for which is described above, had the preferences for its scoring system estimated using the TTO approach [157]. The SF-6D, developed in 2002 by Brazier et al is based on the longer SF-36 measure [153]. It defines health-related quality of life as including physical functioning, role limitations, social functioning, pain, mental health, and vitality and preference weights were derived using standard gamble methodology [153]. The HUI classifies health-related quality of life under 8 headings: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain [155]. The weights for the measure were estimated using standard gamble methods [155,156].

While these measures have some differences, both in what they define health-related quality of life to be and the way they calculated the pre-scored preference weights, they all have an important methodological similarity: the index score is a summary of the value of the health state and not just a description of the health state. As an example of this distinction, take a hip replacement operation. Preference based measures don't just assess whether the hip replacement resulted in increased mobility, but also the value which people get from this increased mobility. This allows comparisons with other interventions on the basis of population preferences and the calculation of QALYs.

2.4. Measuring capability

A primary critique of the extra-welfarist approach is that in practice, the broad evaluative space is significantly limited by a focus on health [7]. Many health and social care interventions may provide benefits other than health [4,66]. For example, a social care intervention to enable elderly people to live in their home may have little impact on the self-reported health of people, but a considerable impact on their independence and enjoyment. Equally a patient who is on an end of life care pathway may value quality time spent with relatives over feelings of good health and vitality. The majority of current health-related quality of life measures fail to assess such outcomes, or such outcomes constitute only a small weight in the overall score. In cases where the benefits of a treatment or an intervention are not captured by the measures used, then intervention may be undervalued. This has led some to advance the measurement of capability as providing a richer evaluative space.

A small number of measures of capability have been developed, generally focusing on the use of questionnaires to assess capability. The process of development of these questionnaires has been varied and attempts occur at different positions along the continuum described by Robeyns et al [111], discussed in Chapter 1. Differences also exist in the influence that the capability approach had on these measures. Some measures were designed with the explicit intent from the outset of measuring capability. Other measures "adopted" the capability approach as a framework for interpreting findings during the methodological process of developing a measure.

This section considers three efforts which have relevance and potential for use in a health or social research setting: the OCAP and OxCAP measures; the ASCOT measure; and the ICECAP measures. Other measurement instruments have been developed, or are in the

process of development, but at the time of writing these are the most advanced and complete measurement instruments.

2.4.1. OCAP and OxCAP measures

The OCAP and later the OxCAP instruments are a group of measure which are developed from the theoretical work of Martha Nussbaum. The development of the OCAP is therefore placed at the theoretical underpinnings end of Robeyns [111] continuum[1]. Anand et al [158] developed a survey which assessed Nussbaum's ten central capabilities using questions asked in the British Household Panel Survey (BHPS), which is a large 5000 house survey designed to be representative of the British population. This measure was then administered to 1048 UK residents in 2005 [159]. The focus of this research was to assess to what extent the attributes included in the OCAP are predictors or covariates of subjective well-being (life satisfaction). Anand et al [159] found roughly a third of the attributes in the OCAP to be predictive of subjective well-being.

Anand notes that using questions which are already in use, such as the ones included in the BHPS is a strength of such an approach [159]. However, this method also has its weaknesses. The use of questions not designed for the purpose of assessing capability may reduce the effectiveness of this measure. In this case, as Anand notes [159], for many of the capabilities being assessed, capability is being inferred from an item which assesses functioning.

Lorgelly et al [113] sought to refine the OCAP developed by Anand and collegues into a measure appropriate for evaluating public health interventions. Using a mixed methods approach, Lorgelly et al [113] reduced the number of items through a reductive process based on the factor loadings from factor analyses and the correlational associations between items.

[1] Robeyns continuum is described in Chapter 1 and defines capability measure development as being either theoretically or methodologically driven.

18 items remained across Nussbaum's ten capabilities. Each question in the OCAP-18 holds equal weight and all responses are coded on a 0 to 1 scale. Consequently, as some of Nussbaum's central capabilities are assessed by more than one question, the OCAP-18 does give extra weight to some of Nussbaum's capabilities [113]. This equal weighting may represent a weakness in the measure, especially when using it for resource allocation decisions.

The most recent development in this research area is the development of the OxCAP-MH by Simon et al [160] through refining the original OCAP for use in a mental health context. Through a process of qualitative discussions with experts, including psychiatrists and psychologists, and pilot work with mental health service users, a 16 item questionnaire was formed. Scores range from 16 to 80, with the intention that having a minimum score different from 0 reflects an ethical standpoint 'that life has its own intrinsic capability value'$_{(p.8)}$[160]. An initial validity assessment has been completed, with correlations found with the EQ-5D-3L [160].

2.4.2. ASCOT

The Adult Social Care Outcomes Tool Kit (ASCOT) began life as the Older People's Utility Scale (OPUS) [161]. The OPUS was designed to accurately reflect the utility of social care interventions. Developed through qualitative work with professionals in the field of social care, five domains were identified as key outcomes of social care: safety, food and nutrition, social participation, personal care, and involvement and control over daily life [161]. Further analysis indicated that the most important of these domains to people receiving social care interventions was personal care. Four further attributes were added in 2005 to make ASCOT

a 9 item questionnaire: activities and occupation, home cleanliness and comfort, anxiety, and dignity and respect [162].

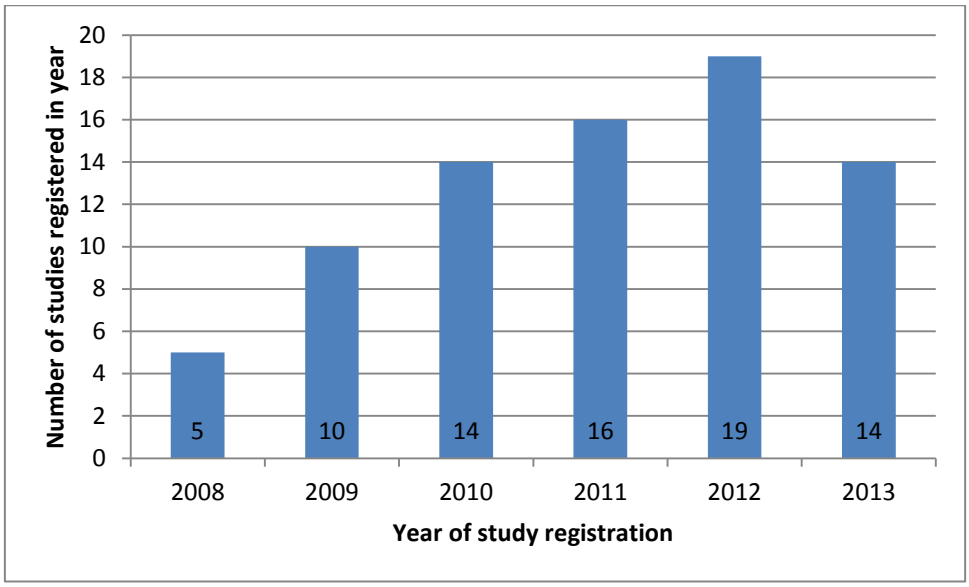
The ASCOT measure seeks to go beyond standard measures of self-reported health and health-related quality of life [163]. Firstly, it recognises that assisting people with disability means going beyond basic functioning, such as sanitation and feeding [163]. It seeks to assess concepts of well-being such as social contact and community status. In recognising this requirement a clear parallel can be drawn with the distinction between well-being and agency defined in the capability approach. Furthermore, the ASCOT measure seeks to assess the impact that a social care intervention has upon whether a functioning is achievable for a person. To do this it makes the distinction between functioning that a person is able to achieve on their own and functioning that they can achieve with the help of the social care intervention. To do this the ASCOT questionnaire assesses whether "my home is as clean and comfortable as I want" rather than "I can clean my home".

While the developers of the ASCOT measure make clear reference to the capability approach, the measure appears to be focused on assessing achieved functioning rather than capabilities. The phrasing of the questions starts with "I have", 2I get" and "I feel" which appear to assess functioning. This tension between the measure and the capability approach is understandable as the theory appears to have been adopted relatively late in the development process, rather than the measure being built on the theory. The capability approach had a greater impact on the later stages of the measure development than the earlier stages. Therefore, the development of this measure can be seen at the methodologically driven research end of the Robeyns continuum [111].

2.4.3. The ICECAP capability measures

The ICECAP-O and ICECAP-A measures are two broad, self-completed well-being measures. Both measures are conceptually linked to Sen's capability approach by defining well-being as the ability to achieve important functionings [164]. They are designed for use in economic evaluation of health and social care. As shown in Figure 1, the number of research projects and randomised controlled trials registered to use the ICECAP-A or ICECAP-O measures has increased year-on-year.

Figure 1: The number of research studies registering to use the ICECAP-A or ICECAP-O measure by year



*Numbers taken from the ICECAP registered users database held at the University of Birmingham

2.4.3.1. ICECAP-O

The development of the ICECAP-O started in 2006, with the work of Grewal et al [109] to identify quality of life attributes of older people for use in a new index measure. A focus of this paper was to go beyond the traditional means through which health economists and researchers have measured quality of life. It was noted by authors that health status, or some limited non-health variables such as material wealth, were often used as proxies for measuring

quality of life [109]. Efforts were made to identify quality of life attributes, rather than health and other attributes that potentially have an effect upon quality of life.

Grewal et al [109] completed 40 in-depth, informant led interviews with members of the British public. Purposive sampling was used to ensure a range of informants with different socio-demographic characteristics were included in the sample. Appropriate qualitative techniques were used to enable informants to fully describe attributes of quality of life, and a thematic framework was used to analyse data. Authors found six broad factors that influence quality of life: activities/doing something; home and surroundings; family and other relationships; standard of health; standard of living/wealth; and religion/faith/spirituality (p1895)[109]. Through these factors the authors identified five attributes which covered the important values from the six factors. These attributes were overarching themes which related to one or more of the factors. For example, feelings of value and self-worth (defined below as "role") were derived through informant's relationships and the activities they were able to do. These five attributes are:

- Attachment which 'incorporates feelings of love, friendship, affection and companionship, sources of which appear to include partners, family, friends, and pets'(p1897) [109].
- Role which 'incorporates the idea of having a purpose that is valued, either by the individual and/or others'(p1897) [109].
- Enjoyment that 'pulls together notions of pleasure and joy, and a sense of satisfaction, sources of which include personal and communal activities'(p1897) [109].
- Security which 'incorporates ideas of feeling safe and secure, not having to worry and not feeling vulnerable'(p1897) [109].

- Control that 'involves being independent and able to make one's own decisions'$_{(p1897)}$ [109].

Authors noted that the ability to function was a 'major theme emerging from the findings'$_{(p1899)}$[109], specifically when informants discussed poor quality of life. Informants discussed poor health, or poor health of a partner, as factors which reduced their ability to achieve the factors and attributes discussed above. The main theoretical work that Grewal et al [109] used to interpret the findings was Sen's capability approach [70]. The major theme of ability to function, which emerged from the research, shows noteworthy similarities to Sen's focus on the importance of what a person is able to do, rather than what they actually do. Furthermore, the finding that health is not simply an end in itself, and that people may value factors other than health, sits neatly with Sen's distinction between well-being and agency.

The five attributes of Attachment, Role, Enjoyment, Security and Control were used by Coast and colleagues [100,118] in the development of the ICECAP-O instrument. The conceptual terms were refined into a classification system that would hold meaning for the general population of 65 and over. For example, the term "role" was rephrased to "doing things that make me feel valued". This was done through semi-structured interviews with 19 informants. Iterative techniques were used to refine and test language for both the attributes themselves and the response levels for each attribute. The result of this development process was five attribute items, each with four response levels (see Appendix 1).

The terminology of the ICECAP-O measures was designed to fit with the capability approach. Questions were prefixed with "I can" and "I am able" and are designed to assess people's ability to function rather than their actual functioning. Coast et al [118] note that while the capability approach is incorporated into the measure, the measure is also highly influenced by

health economic norms of assessing health through people's perceptions, therefore not offering an objective assessment of capabilities. The authors note 'that "pure" capabilities researchers would refer to this instrument as an index of "perceived capability"'$_{(p.881)}$[118].

Coast et al [118] completed the valuation of the ICECAP-O measure using a type of discrete choice experiment, the best-worst scaling method. This is a scenario-based procedure whereby respondents are asked to choose the best and worst scenarios from a selection of methods [165]. From these choices values for the capability sets were derived [118]. This valuation considered attributes from full capability to no capability and did not make any assumptions about where death fell on this scale [118]. Having no control was associated with the lowest level of value for an attribute. Differences were seen in the value that participants placed on each level of each attribute. Changes between levels at the top of the item were of less value to respondents than changes between levels at the bottom of the measure. For example, the difference between "no" and "a little" Attachment was considerable and larger than the difference for other items. The difference between "all" and "a lot" of Security was larger than the differences between the top two levels on other items.

2.4.3.2. ICECAP-A

The development and valuation of a capability measure for the whole of the adult population began with work by Al-Janabi et al [115] to identify and refine the attributes for the ICECAP-A instrument. This two stage work first involved 36 semi-structured interviews with a purposively selected sample of the general population. The focus of this first stage of work was to identify what was important to people's lives. The analysis identified a similar, but not identical, list of five attributes: Stability, Attachment, Autonomy, Achievement and Enjoyment. 'Attachment, Autonomy and Enjoyment are almost identical, albeit with some

adjustment in wording, to three attributes in the ICECAP-O measure'$_{(p173)}$[115]. The Stability item has similarities with the Security item in the ICECAP-O, but with a greater focus on the present [115]. It refers to a desire for continuity in life and is affected by a broad range of factors including health, employment status and fear of crime [115]. The Achievement item shows similarities to Role, but goes beyond it in scope. It refers to the ability of an individual to move forward in life, achieve goals and is associated with success at work, having a family and owning things [115].

The second stage of this work was to establish the appropriate terminology to be used in the instrument [115]. Based on 18 semi-structured interviews the terms Stability, Attachment and Autonomy were judged to be unsuitable for inclusion, and the terminology was changed: Stability became "settled and secure"; Attachment became "love, friendship and support"; and Autonomy became "independence"[115]. The Achievement item was supplemented with "progress" and Enjoyment was supplemented with "pleasure", to increase understanding [115]. As with the ICECAP-O, questions were prefixed with "I can" and "I am able" in order to assess capability rather than achieved functioning. For the final version of the ICECAP-A see Appendix 2.

The best-worst scaling method was used to estimate capability values [166]. 413 individuals were sampled through a random sampling method stratified by geographic area of the UK and socio-economic deprivation [166]. The results indicated that people held strong preferences for Stability and Attachment attributes, with Autonomy being the next most important item. As a result, Attachment and Stability each account for 22% of the weight in the estimated tariff score, with other items accounting for roughly 18% weight each [166]. As with the ICECAP-O, there were differences in the values that participants placed on the levels of each

item. Again, this results in the situation where change at the top of the measure (between the top two levels of an item) is valued less than changes at the lower end of the measure.

2.4.4. Capability measures and public policy in the UK

In 2013 the National Institute for Health and Care Excellence (NICE) was given responsibility for developing guidance and standards for social care in England [167]. NICE discharges this responsibility in co-operation with the Social Care Institute for Excellence (SCIE), which leads the NICE Collaborating Centre for Social Care (NCCSC) [168]. Both NICE and SCIE have indicated that when measuring the effectiveness of social care interventions a flexible approach is needed which reflects 'the nature of effects delivered by different social care interventions or programmes' [169]. Both organisations have recognised the need for the use of broad preference based measures, with the ICECAP measures and the ASCOT measure highlighted as appropriate for use [169,170]. SCIE has noted that both the ASCOT and ICECAP measures are 'relatively new and validity and reliability are still being tested'[170]. While, the psychometric properties of these measures are being determined NICE and SCIE are currently encouraging the use of the ASCOT and ICECAP measures in tandem to understand how they complement each other.

2.4.5. Thesis and the ICECAP-O and ICECAP-A

This thesis seeks to assess the validity and responsiveness of the ICECAP-O and ICECAP-A capability measures in a trial setting. The ICECAP-SCM is an additional measure from the "ICECAP family" designed for use with patients at the end of life. It is a supportive care measure, which is not designed for use as a quality of life measure and therefore not considered in this thesis. There are differences between the ICECAP-A and ICECAP-O in the attributes which are assessed. Where the same attribute is assessed in both measures, there

are differences in the wording of the items. Validity results that are found in one measure cannot blindly be applied to the other. Therefore, in the methodological review and the primary research in this thesis the results pertaining to the ICECAP-A and ICECAP-O are clearly demarcated. In the discussion this distinction is somewhat relaxed, with references to "ICECAP measures" where appropriate. However, when a discussion point refers only to one measure this is clearly stated.

2.5. A methodological review of the psychometric properties of the ICECAP measures

2.5.1 Aim of review

The aim of this review was three-fold. The first aim was to identify, compile and document the body of research providing information on the validity of the ICECAP-O and ICECAP-A measures. A small body of validation research has quickly developed since the development of the ICECAP-O measure in 2008. Identification of this existing research is useful in informing future directions for research. The second aim was to synthesise this research to assess the psychometric properties of the ICECAP-A and ICECAP-O measures. The ICECAP measures are two relatively new measures and validation work is at an early stage. Therefore, it was anticipated that in many circumstances it would not be possible to form firm conclusions on validity from the evidence found. In these circumstances an effort was made to fully describe the results. A third aim was to inform the development of evidence-based hypotheses which were used in the validation analyses presented in Chapter 7 of this thesis. As described in Chapter 3, the formation of evidence-based hypotheses is an important step in a rigorous and scientific approach to validation. The hypotheses formed are reported in the methods chapter rather than here.

2.5.2. Review methodology

The review methodology was specified and documented in advance of the systematic review.

2.5.2.1. PICOS criteria

The PICOS (Population, Intervention, Comparators, Outcomes, Study Design) criteria, from the Centre for Reviews and Dissemination, University of York, are a useful way of framing the question which a review seeks to answer [171] and allowing clear communication of the research that is to be included in a review. The PICOS criteria for this review were formed with the knowledge that a small body of literature existed. The population, interventions and outcomes that were eligible for inclusion in the review were kept deliberately broad. Had a greater body of literature existed the review could have focused on validity in select populations or circumstances, such as with patients or in randomised controlled trials, but this was not possible.

2.5.2.1.1. Population

This review did not exclude any studies based on the characteristics of the population used. General population and patient populations were included.

2.5.2.1.2. Interventions

This review included interventional and non-interventional primary research.

2.5.2.1.3. Comparators

Comparators normally refer to the control treatment used in a comparative study such as a randomised controlled trial. In the context of validity research, it should be considered as referring to constructs or anchors against which a measure is validated. No study will be excluded on the basis of the comparators used.

2.5.2.1.4. Outcome

Any outcome which provided information on the psychometric properties of either the ICECAP-A or ICECAP-O was included. The majority of outcomes are presented as descriptive statistics, correlation analysis, univariate statistics such as chi-squared or ANOVAs or multi-variate statistics such as regression models.

2.5.2.1.5. Study design

Qualitative and quantitative study designs were included in the review.

2.5.2.2. Review search strategy

The main search strategy was completed in July, 2013 and was updated in November, 2013.

A 3 stage search strategy was employed:

1. A keyword/MeSH term search used the OVID search facility. The OVID search facility provides an online platform through which scientific research can be systematically accessed. An initial scoping search was used to identify appropriate search terms. Based on the results of this scoping search the terms "ICECAP*" and "Quality of life" and the terms "ICECAP*" and "capabilit*" were searched on all databases available through the OVID search facility (these included Embase and Medline). Placing a "*" after the words expanded the terms so that a broader range of words or terms were searched.
2. A comprehensive key person search was completed via e-mail correspondence. The database of registered ICECAP users held at the Health Economics Unit, the University of Birmingham and a list of those attending the 2011 ICECAP workshop were used to identify people who had potentially assessed the psychometric properties

of the measure. All individuals providing an email address were contacted and asked two questions: 1) had they completed any validation work on the ICECAP measures; and 2) were they aware of anyone who had (see Appendix 3).

3. A forward/backward citation search was completed on all articles included from steps 1 and 2 of the search strategy. The forward search was completed on OVID, while the backward search was completed via a manual inspection of the article reference lists.

2.5.2.3. Article screening

A two-step article screening process was employed:

1. Title and abstract screen. All articles identified had their titles and abstracts screened for relevance. In the case of uninformative titles and abstracts not allowing understanding of content, the article was forwarded into the full text review. A second assessor check was completed on all identified references.
2. Full text review. A full text review of articles was completed to check for relevance.

2.5.2.4. Inclusion and exclusion criteria

Inclusion criteria were kept deliberately broad to allow identification of all relevant studies in this relatively new and under researched area. Articles were selected for inclusion if they met the following criterion:

- The study provided empirical data on the psychometric properties of the ICECAP-O or ICECAP-A measure.

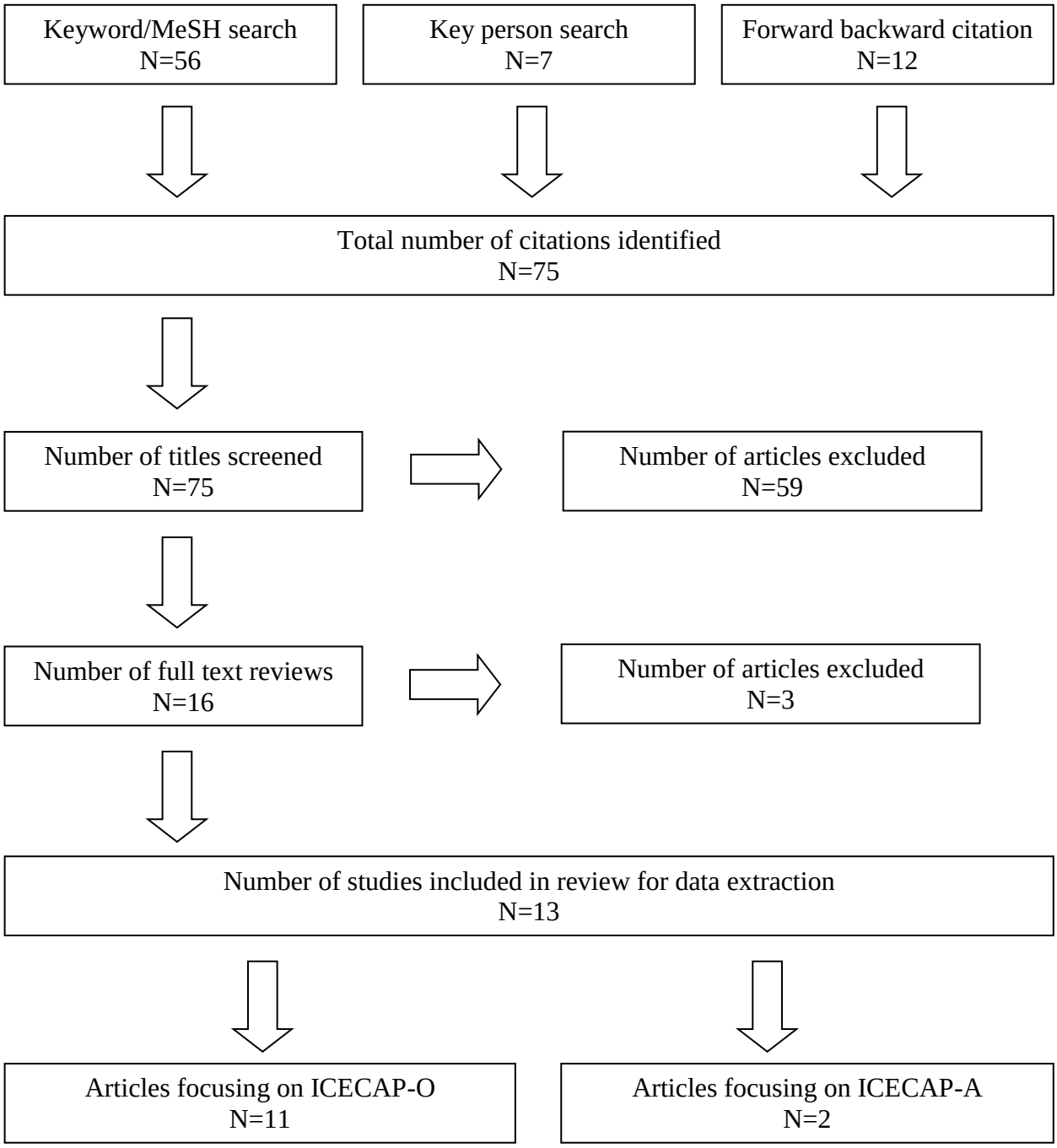
Articles were excluded from the review if they met one or more of the following criteria:

- The study presented theoretical debate, without empirical data.
- The article was not written in the English language.

2.5.3. Search results

The keyword/MeSH term searched identified 56 references of potential relevance to the review. The database of registered ICECAP measure users and the list of attendees at the 2011 ICECAP workshop identified 59 "key people" with an e-mail address. E-mails were sent to all these people. Responses were received from 36 people, with the majority stating they did not know of any relevant papers. Seven potentially useful references were identified through these key persons. The forward/backward citation search identified 12 citations. Of the 75 identified articles 59 were excluded after the title/abstract screen. 3 were excluded after the full text review, leaving 13 articles for inclusion in the review. Results are reported in line with PRISMA guidelines.

Figure 2: Methodological review citation identification, inclusion and exclusion



2.5.3.1. Study characteristics of articles included in review

Thirteen studies [114,116,117,172,173,173–182] providing information on the psychometric properties of the ICECAP measures were identified and included in the review. Table 1 provides a summary of the study characteristics, while Appendix 4 provides an in-depth look at the characteristics of each study.

Eleven studies provided information on the ICECAP-O measure [117,172–180,182], while two provided information on the ICECAP-A measure [116,181]. The predominance of papers referring to the ICECAP-O is likely due to the greater length of time for which this measure has been in use compared to the ICECAP-A. The majority of the studies were small to medium sized quantitative works, with two qualitative studies identified. General population and patient samples were used. Studies collected data on a number of comparator measures; the most common of these was the EQ-5D-3L, administered in the majority of studies. Socio-demographic, disability and well-being data were also frequently collected. The 13 articles identified were produced by four research groups: Coast and colleagues, based at the University of Birmingham, UK; Ratcliffe and colleagues, the University of Adelaide and Flinders University, Australia; Makai and colleagues, Erasmus University and Davis and colleagues, the University of British Columbia

The majority of studies were cross-sectional studies, using data from one time point. The characteristics of the samples used in the studies differed (see Appendix 4). The predominance of studies assessing the validity of the ICECAP-O means that (appropriately) the studies have used elderly populations. The majority of studies have sampled more female participants than male. Studies have used both the general public and patients as participants. Those studies which used patient populations held some similarities. Patients were generally

non-hospitalised patients: either rehabilitation patients or falls prevention patients. The only study to use hospitalised patients was Makai et al [172]. The average health state of participants in these studies, judged by mean scores on the EQ-5D-3L, varied from good to very poor.

Studies used a number of statistical approaches: descriptive statistics, multiple bivariate tests and multivariate tests. Some studies used the UK ICECAP-O measure or/and algorithm in non-UK populations. The studies by Davis [173,174] used the same data set of patients visiting a falls prevention clinic. The same population of post-acute hospital rehabilitation patients was used in studies by Ratcliffe [176], Couzner [175] and Couzner [178].

Table 1: Characteristics of research articles included in review

Characteristic	Number of studies
ICECAP version	
• ICECAP-O	11
• ICECAP-A	2
Study type	
• Quantitative	11
• Qualitative	2
Size of study sample (of quantitative studies n=11)	
• Small (<200)	3
• Medium (200-500)	5
• Large (>500)	3
General population or patients*	
• General population	6
• Patients	8
Country	
• Australia	4
• Canada	2
• Netherlands	2
• United Kingdom	5
Comparators used**	
• EQ-5D-3L	10
• Physical functioning measure (not EQ-5D-3L)	3
• Mental health measures	2
• Measures of well-being, happiness or life-satisfaction	4
• Socio-demographic characteristics	5
Research group	
• Coast and colleagues (Birmingham and Bristol)	5
• Davis and colleagues (British Columbia)	2
• Makai and colleagues (Erasmus)	2
• Ratcliffe and colleagues (Adelaide and Flinders)	4

* One study used patients and general population; **Multiple studies used multiple comparators.

2.5.4. ICECAP-O results

2.5.4.1. ICECAP-O completion rates

Completion rates for the ICECAP-O measure were assessed by Davis et al [173]; Makai et al [172]; Coast et al [117]; Couzner et al [175]; Flynn et al [177]; Makai et al [180]and varied between 92% and 98%. No conclusion of how the completion method (self-completed or administered) or setting (clinical or home) affects the results could be made from this small number of studies.

Qualitative think aloud work by Horwood et al [182] showed that informants completing the measure showed a low level of struggle and minimal difficulties. 20 participants completed the measure while concurrently verbalising their thoughts. Results showed that, of the 100 items completed (i.e. 5 items completed by 20 informants), informants struggled on 7% of these. The majority of problems were caused by informants struggling with comprehension of the item.

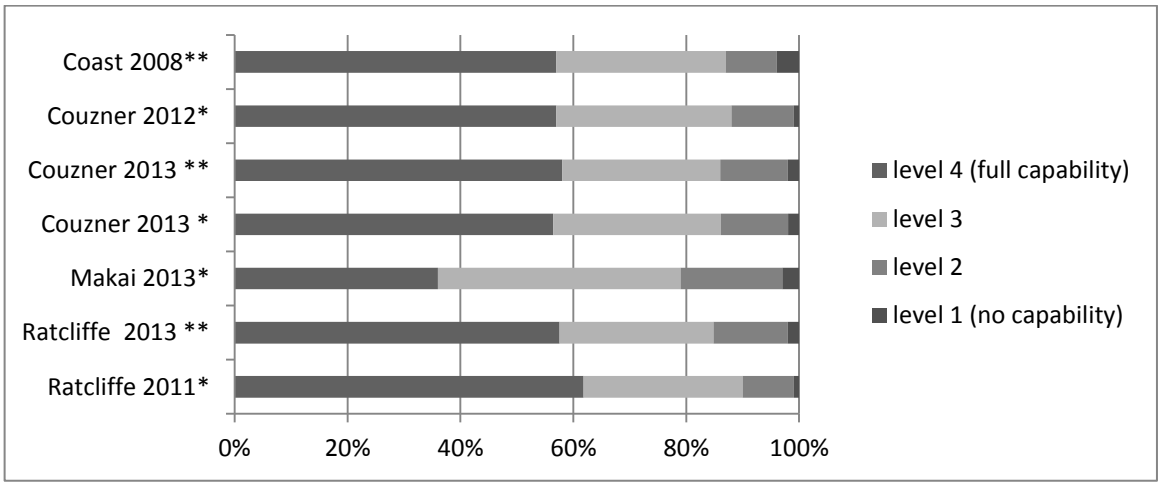
2.5.4.2. Response distribution

The distribution of responses on the ICECAP-O tariff score or ICECAP-O individual items was assessed by a number of the articles identified. Two studies assessed the distribution of ICECAP tariff scores (Flynn et al [177]; Davis et al [173]). Both studies showed a clustering of values above 0.8, with a very small number of responses under 0.6. Coast et al [117], Makai et al [180], Ratcliffe et al [176], Couzner et al [178], Ratcliffe et al [179] and Couzner et al [175] assessed the spread of responses across levels of each item of the ICECAP measures, discussed in relation to each attribute below.

2.5.4.2.1. Attachment response distribution

Studies showed similar response distributions for the Attachment item. Between 55% and 60% of respondents categorised themselves as having full capability and sequentially lower percentages for the lower levels (see Figure 3). The study by Makai and colleagues was an exception to this pattern, with a higher percentage of patients recently discharged from hospital categorising themselves in the second level of capability. Authors indicate that this difference may be due to worse mobility in their population than other study populations and general populations [180].

Figure 3: Response distribution for Attachment item in research articles included in review



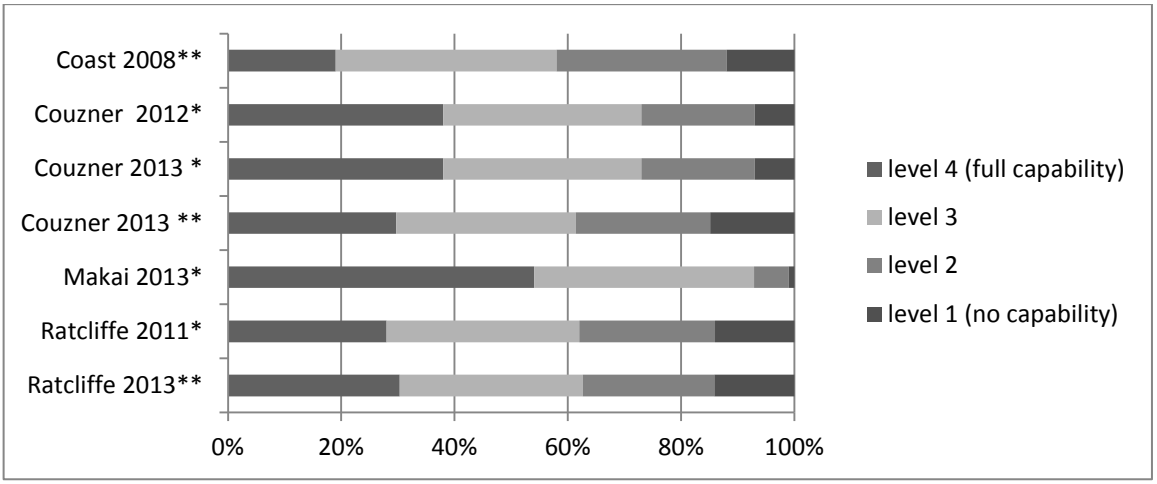
*patient values, **general population values

2.5.4.2.2. Security response distribution

In comparison to the Attachment item, a lower percentage of respondents categorised themselves as having full capability for the Security item and a greater variability in responses was found (see Figure 4). Reponses were more evenly split between the top two levels. Between 1% and 15% of respondents categorised themselves as having no capability on this item. A different response pattern can be seen in Makai et al's study, which showed a higher percentage of respondents reported full capability for Security. Authors hypothesise that this

difference may be due to cultural differences in the Dutch population. This is the second study to report high values for the ICECAP-O Security item in the Dutch population, with Makai et al [172] reporting high values in this item (this was not reported in these figures as the value was not expressed as a percentage of respondents answering each level, rather an unweighted mean and median score).

Figure 4: Response distribution for Security item in research articles included in review

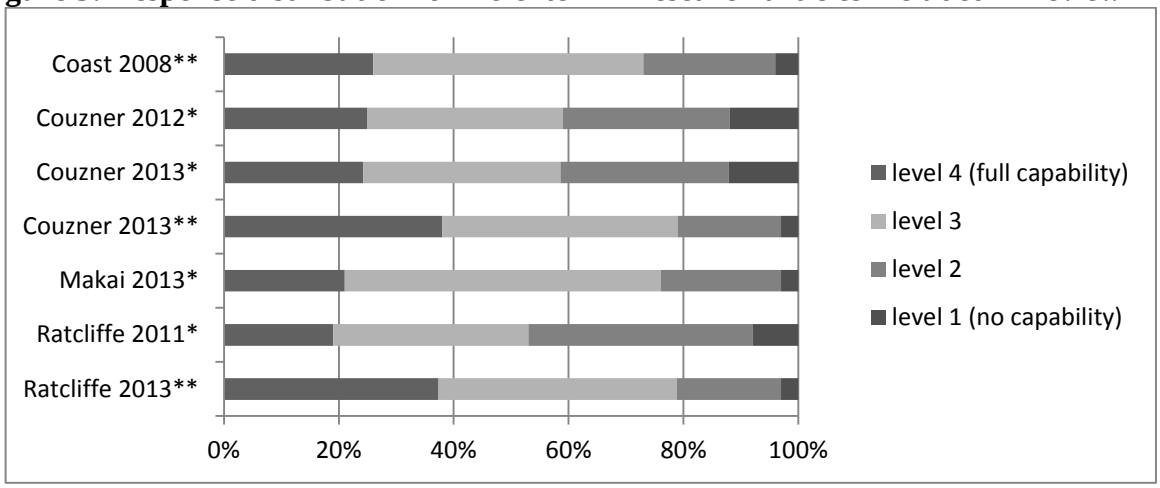


*Patient values, **general population values

2.5.4.2.3. Role response distribution

The most frequent response for the Role item was level 3, with percentages varying between 34 and 55% (see Figure 5). The percentage of respondents categorising themselves as having full capability varied around 20% to 25%, with exceptions from Couzner et al's 2013 study general population values and Ratcliffe et al's 2013 study which found close to 40% with full capability.

Figure 5: Response distribution for Role item in research articles included in review

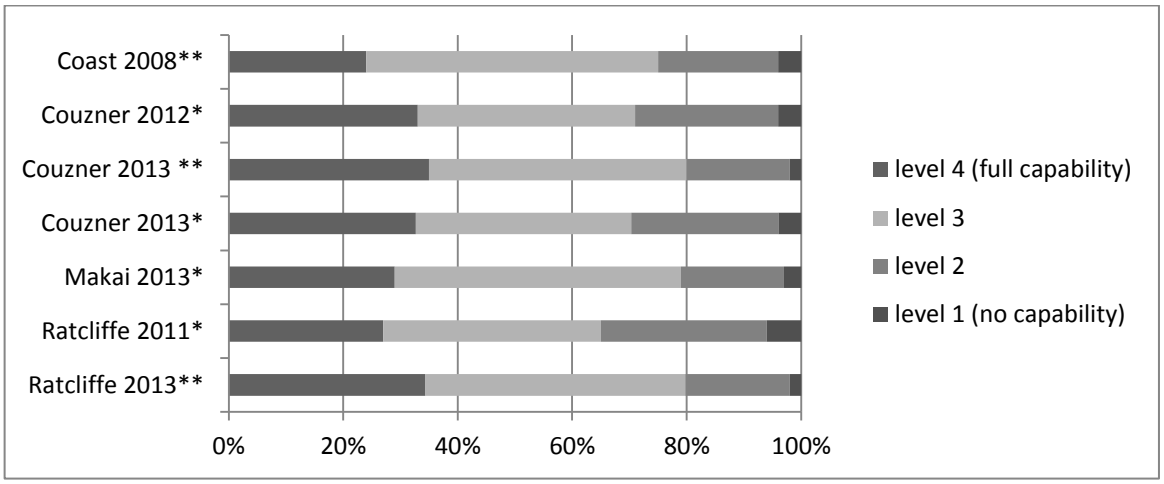


*patient values, **general population values

2.5.4.2.4. Enjoyment response distribution

The response pattern to the Enjoyment item showed a (somewhat) even distribution of responses between the top three capability levels (see Figure 6). A small number of respondents categorised themselves as having no capability.

Figure 6: Response distribution for Enjoyment item in research articles included in review



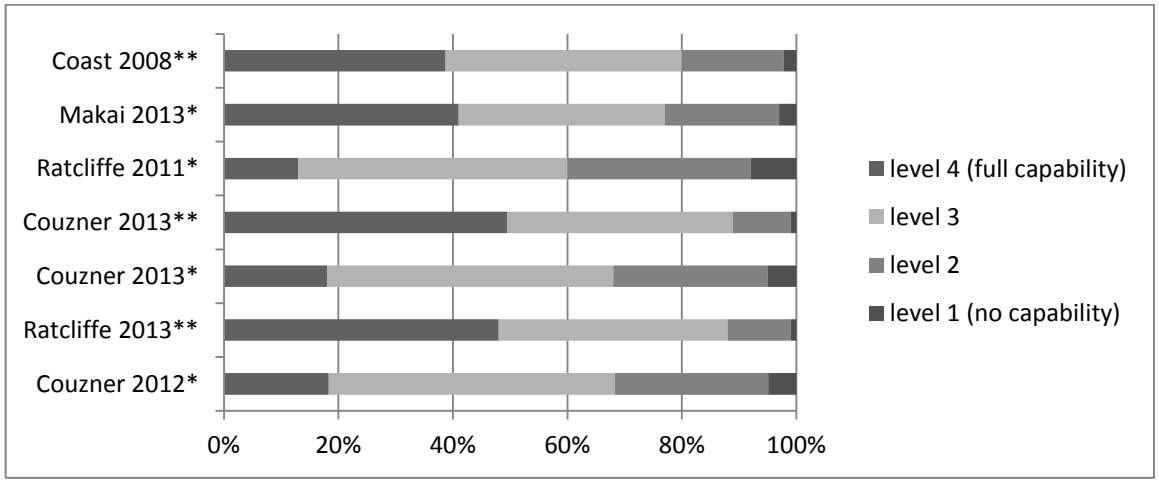
*patient values, ** general population values

2.5.4.2.5. Control response distribution

Notable variability exists in the response distributions of the Control items. A division is apparent between studies: three studies found that below 20% categorised themselves as

having full-capability and four studies found that roughly 40% categorised themselves as having full capability (see Figure 7). Most of the studies which found the percentage of participants in full capability approaching or over 40% were using general population samples and data were collected in a non-medical setting. All other studies used patient populations.

Figure 7: Response distribution for the Control item in research articles included in review



*patient values, **general population values

2.5.4.3. ICECAP-O measure and socio-demography

A number of socio-demographic variables were used in studies identified. There was little uniformity in the comparators used and the reporting methods, which limits the ability of a methodological review to synthesise results [183]. Three topics were identified where multiple studies had reported the same (or very similar) outcomes and it was possible to synthesise these in a narrative review. A detailed review of socio-demographic variables, as well as other variables used in studies, is presented in Table 2.

2.5.4.3.1. Gender

No study hypothesised or found any association between the gender of participants and the ICECAP-O tariff score or scores on individual items.

2.5.4.3.2. Age

Studies that provided hypotheses expected increasing age to be associated with decreasing ICECAP-O score. Mixed results for the association between ICECAP-O scores and age were found. Makai et al [180] and Flynn et al [177] found associations between the age of the participant and the ICECAP-O tariff score, with younger individuals having higher capability. Coast et al [117] showed associations between age and the items of Role, Enjoyment and Control. These studies contrast with studies by Ratcliffe et al [176] and Couzner et al [175] which found no association between the age and ICECAP-O items and Couzner et al [175] which found no association with the overall ICECAP-O score.

2.5.4.3.3. Relationships and living with partner

Studies hypothesised that higher ICECAP-O capability scores would be expected in those living with a partner or in a relationship. The ICECAP-O items of Attachment [117,175] and Role [117] were associated with a person's relationship or marital status, with those in relationships reporting more capability. Flynn et al [177] found living alone to be associated with ICECAP-O tariff scores, with those living independently having lower scores.

Table 2: Summary of primary hypotheses, findings and conclusions from research articles referring to ICECAP-O included in review

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Coast (2008)	*Socio-demographic* Weak evidence that age would be associated with Role and Control, but not with other items. Sex would not be associated with any item. *General well-being* Assessment of general well-being would be associated with a number of items. *Contact with others* Strong evidence that Attachment and weaker evidence that Security, Enjoyment and Role would be associated with contact with others. *Health* Assessments of health would show associations with Control, Role and Enjoyment, weaker evidence of associations with Security and no association with Attachment. *Social support* Strong evidence of associations between measures of social support and Security, Role, Enjoyment and Control.	*Socio-demographic* Strong evidence of associations between age and Control and Role, and weak evidence of association with Enjoyment. No relationships between sex and any item. *General well-being* Evidence of general well-being associated with items. *Contact with others* Marital status was associated with items as anticipated. Contact with others unexpectedly showed weak or no association with Attachment or Security. *Health* Strong evidence of association of health with all items. Mental health most strongly associated with Attachment and Enjoyment; physical health with Role, Enjoyment and Control. *Social support* Receipt of informal care highly associated with Enjoyment, Control, Role and Security.	In general relationships that were anticipated *a priori* were found. This indicates that the ICECAP-O measure is measuring what it is designed to measure.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Couzner (2012)	*Health* Anticipated that Control, Enjoyment and Role would be associated with health status measured by the EQ-5D-3L. It was anticipated that EQ-5D-3L items of mobility, self-care and usual activities would be associated with ICECAP-O scores. *Quality of care transition* Those with higher quality of care transitions, measured by the CTM-3 would report higher ICECAP-O scores. *Medical care received* It was expected that those receiving outpatient rehabilitation would report higher ICECAP scores than inpatients.	*Health* ICECAP-O tariff scores were correlated with EQ-5D-3L index scores (0.437). Control, but not Enjoyment and Role, were associated with EQ-5D-3L scores. Mobility, self-care, usual activities and anxiety and depression were associated with EQ-5D-3L index scores. *Quality of care transition* CTM-3 scores were significantly correlated with ICECAP-O tariff scores *Medical care received* Results not found *Socio-demographic* Age was not associated with any item of the ICECAP-O. Whether the participant had an informal carer was associated with the Role item.	Relationships between ICECAP scores and self-reported health and quality of care transitions suggest that these may influence some, but not all aspects of a person's capability. These factors are more influential than socio-demographic factors on capability scores.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Couzner (2013)	*Patient vs general public* It was anticipated that the patient sample would report lower ICECAP-O scores than the general population sample. The magnitude of the sample differences on the EQ-5D-3L and ICECAP-O would likely differ due to different theoretical underpinnings of the measures.	*Patient vs general public* The general population reported higher ICECAP-O scores that the patient population. This difference was more pronounced for younger patients. Larger, more pronounced, differences were found between patients in EQ-5D-3L scores than in ICECAP scores.	No conclusion referring directly to the validity of the ICECAP-O was given, however, author emphasised the importance of the finding that the difference between patients and general population in EQ-5D-3L scores are more pronounced than the difference in ICECAP scores.
Davis (2012)	*EQ-5D-3L* Usual activities would show association with the Role item of the ICECAP-O. Self-care would demonstrate agreement with Control.	*EQ-5D-3L* The correlation between the ICECAP-O and the EQ-5D-3L was 0.47. Significant differences between the self-care and Control and Role and usual activities were seen. A factor analysis indicated that a two factor solution was the best fitting model, with the majority of ICECAP-O items in one factor and the majority of EQ-5D-3L items in another.	Results indicate that the ICECAP-O and the EQ-5D-3L provide "largely unique and complementary information".

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Davis (2012)	No-hypotheses were provided.	*EQ-5D-3L* The ICECAP-O and the EQ-5D-3L was significantly correlated at 0.47. *Instrumental activities of daily living* Role, Enjoyment and Control were significantly correlated with IADL. *Short Physical Performance Battery* Role and Control were significantly correlated with the SPPB *Physiological Profile Assessment* Control was significantly correlated with PPA. *Mini-Mental State Exam* Security was significantly correlated with mini mental state exam (MMSE).	In an older adult population both measures (EQ-5D-3L and IECCAP-O) provide valuable and unique information.
Flynn (2011)	Relationships were anticipated with: health, cohabitation, age, qualifications and receipt of benefits.	*Socio-demographic* No difference in ICECAP-O tariff scores was found by gender or ethnic group. Younger, qualified, co-habiting or those not receiving benefits reported higher ICECAP scores than older, non-qualified, those living alone or those receiving benefits. *Health* Significant differences in ICECAP-O scores were seen between those reporting good, fairly good and not good general health.	Results provided support for the validity of the ICECAP-O tariff scores. Health is not the only factor explaining ICECAP-O scores, but it is clearly an important one.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Horwood (2013)	Qualitative work – no hypotheses provided	*Comprehension problems* 5 informants were identified as having problems completing one or more items of the measure, 9 informants struggled with one or more item but answered appropriately, 6 participants had no problem. Of the 100 item responses analysed 7% of them showed problems, with comprehension problems being the most common. *Item-by-item* Informants queried whether the Attachment item was relevant to their disease. Informants questioned both the time-frame and the focus of the Security item.	The ICECAP-O measure performed well in this clinical population, with only a small number (7%) of responses showing struggle when completing items. Results indicate that the measure may benefit from an introductory statement explaining that general quality of life is being assessed, and individual items may benefit from a short statement clarifying what the measure is trying to assess.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Makai (2012)	*Restrained vs non-restrained patients* Differences in ICECAP-O scores between psycho-geriatric patients who are restrained and those who are not. *Overall life satisfaction and QoL* Overall measures of life satisfaction and QoL would be associated with ICECAP-O scores.	*Restrained vs non-restrained patients* Significant differences between the two groups of patients in all ICECAP items apart from Security. *Overall life satisfaction and QoL* ICECAP-O scores significantly correlated with Cantril's ladder and overall life satisfaction question. *Health and EQ-5D-3L* The EQ-5D-3L index and EQ-5D-3L VAS score was significantly correlated with ICECAP-O. The strength of the correlation was between 0.43 and 0.57 depending on the version of the EQ-5D-3L (family or nursing).	Study showed reasonable convergent and divergent validity evidence for the ICECAP-O measure. The measure seems a "promising" instrument for use. Moderate strength of the correlation between the EQ-5D-3L and the ICECAP should be considered in light of the broader evaluative space of the ICECAP measure.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Makai (2013)	*Overall well-being and QoL* ICECAP-O was anticipated to correlate more strongly with Cantril's ladder and SPF-IL than measures of health. *Health* The ICECAP-O was expected to correlate with the EQ-5D-3L, IADL, GDS and SF-20, but the correlation will not be as strong as with measure of well-being.	*Overall well-being and QoL* ICECAP-O tariff scores were strongly correlated with Cantril's ladder and moderately correlated with SPF-IL. *Health* The EQ-5D-3L index score was moderately correlated with the ICECAP-O tariff score. Correlations with other health measures were moderate to weak. ICECAP-O was able to discriminate between multi-morbid and single-morbid patients, between depressed and non-depressed and between IADL dependent and non-dependent people. *Socio-demographic* The ICECAP-O was able to discriminate between "young-old" and "old-old" patients.	The ICECAP-O measure showed good convergent validity with measures of health and well-being. The correlations with health were unexpectedly similar to those of well-being. The ICECAP-O seems to be a promising instrument for use.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Ratcliffe (2011)	*Health* It was anticipated that there would be strong relationships between ICECAP-O tariff scores and EQ-5D-3L index scores. Mobility, self-care and usual activities were expected to be strongly associated with ICECAP scores. Control, Role and Enjoyment were hypothesised to be more likely to associate with EQ-5D-3L scores than Security and Attachment. ICECAP-O scores would be inversely related to scores of the Modified Rankin Scale. *Hope* Higher scores on the Herth Hope Index would be associated with higher ICECAP-O scores. *Quality of care transition* It was anticipated that there might be an association between the CTM-S and the ICECAP-O tariff scores.	*Health* A correlation of 0.418 was seen between the ICECAP-O and the EQ-5D-3L. EQ-5D-3L index scores were associated with all ICECAP items apart from Role. The EQ-5D-3L item of usual activities was associated with Role, Enjoyment and Control at the 5% level. A negative correlation of -0.286 was seen with the Modified Rankin Scale. *Hope* The Herth Hope Index showed a 0.402 correlation with the ICECAP-O tariff score and was associated with Security at the 5% level. *Quality of care transition* A correlation of 0.259 was seen between the CTM-3 and ICECAP-O tariff score.	The strong empirical relationships between the comparator measures in this study and the ICECAP-O support the construct validity of the measure in a clinical rehabilitation study.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures
Ratcliffe (2013)	No hypotheses were provided.	*Socio-demographic* Carers and non-carers had similar ICECAP-O tariff scores, with a small indication that carers may have higher scores. Carers were more likely than non-carers to report the highest capability of Role and Control. Younger, native Australian and higher income earners reported higher ICECAP-O tariff scores than their counterparts.	No conclusion referring directly to the validity of the ICECAP-O was provided by authors.

2.5.4.4. ICECAP-O measure and health

The majority of the articles assessed the relationship between ICECAP-O measures and health through the use of patient-reported outcome measures, such as the EQ-5D-3L. These results are summarised below.

2.5.4.4.1. Objective measures of health

Couzner et al [178] used an objective comparison between post-acute care patients and the general public. ICECAP-O scores in the general public were higher than in the patient population (0.795 against 0.753). This difference appears to be less pronounced than the difference in EQ-5D-3L scores between the two groups (0.789 against 0.595). Effect sizes which would increase the certainty of this comparison were not provided by authors.

2.5.4.4.2. EQ-5D-3L

The most frequently administered health measure in these studies was the EQ-5D-3L. Studies that provided hypotheses all expected moderate correlations or associations between the two measures. Studies found moderate, positive and statistically significant correlations between the EQ-5D-3L index score and the ICECAP-O tariff score [172,173,175,176,180], shown in Table 3.

Table 3: Correlations between EQ-5D-3L and ICECAP-O in research articles included in review

	Davis, 2012	**Makai, 2012**	**Couzner, 2012**	**Ratcliffe, 2011**	**Makai, 2013**
Correlation between ICECAP-O tariff score and EQ-5D-3L index.	0.47	0.43 (family) 0.57 (nursing)	0.44	0.42	0.40

Analyses showed that higher EQ-5D-3L index values were recorded in participants categorising themselves as having higher levels of capability on individual ICECAP items [117,175]. Enjoyment and Control items were strongly associated with the EQ-5D-3L index, while the Attachment item was not associated [117,176,180]. A factor analysis of the EQ-5D-3L and the ICECAP-O completed by Davis et al [173] showed a two factor solution to be the best fitting solution: one factor, termed by the authors "physical functioning", including predominantly EQ-5D-3L items and one factor "psychosocial well-being", including all the ICECAP-O items and the anxiety and depression question from the EQ-5D-3L.

2.5.4.4.3. Measures of general health

Two studies included brief, general assessments of health. These assessments used single question measures. Flynn et al [177] found self-reported general health to be significantly associated with ICECAP-O scores, with those reporting better health also reporting higher ICECAP-O scores. Coast et al [117] found general health to be significantly associated with the items of Role, Enjoyment and Control at the 1% significance level.

2.5.4.4.4. Measures of physical ability and disability

Studies that provided hypotheses anticipated that measures of disability would be inversely associated with ICECAP-O scores (higher disability, lower ICECAP scores). Fours studies compared ICECAP-O scores with measures of physical ability or disability [117,172,174,176]. Two studies used the Instrumental Activities of Daily Living scale as a comparator. Makai et al [172] showed a significant moderate correlation of 0.51 with the ICECAP tariff scores, while Davis et al [174] did not find the measures to be significantly correlated. These differences should be viewed in light of further results from Davis et al

[174] which showed the Physiological Profile Assessment (PPA) and Short Physical Performance Battery (two measures of physical ability) not to be significantly correlated with the ICECAP-O measure. Ratcliffe et al [176] showed a weak correlation between the Modified Rankin Scale and the ICECAP-O measure. Studies found Role and Control were strongly associated with measures of disability [117,172,174,176] and Coast et al [117] also found Enjoyment to be associated.

2.5.4.4.5. Measure of psychological health

Mixed evidence was found from the three studies which used measures of psychological health in their studies [172,176,180]. Makai et al [172] showed non-significant associations between the HADS and the ICECAP in both nurse and family completed measures. In a post-hospitalised population the Geriatric Depression Scale (GDS) was strongly correlated with the ICECAP-O measure [180]. The Herth Hope Index (HHI) showed moderate correlations with the ICECAP-O measure [176].

Three studies assessed the associations between the anxiety/depression item of the EQ-5D-3L and ICECAP-O items [117,175,176]. Mixed and conflicting results were found. Coast et al found the item to be associated with all items apart from Control. Couzner et al [175] found the item to be associated only with Control, while Ratcliffe et al's [176] results indicate an associate with Security, Role and Enjoyment.

2.5.4.5. ICECAP measures and well-being

Strong evidence was found that well-being is significantly associated with ICECAP scores. Makai et al [172] found two assessments of well-being, Cantrill's Ladder and the SPF_IL, to be significantly associated with the items of Role, Enjoyment and Attachment and to have

moderate to strong correlations with the overall ICECAP tariff scores. Makai et al [180] found Cantril's ladder and an overall life satisfaction measure to be associated with ICECAP-O scores. Using a simple, one question assessment of well-being Coast et al [117] found all ICECAP items to be associated with well-being.

2.5.5. ICECAP-A

2.5.5.1. ICECAP-A completion rates

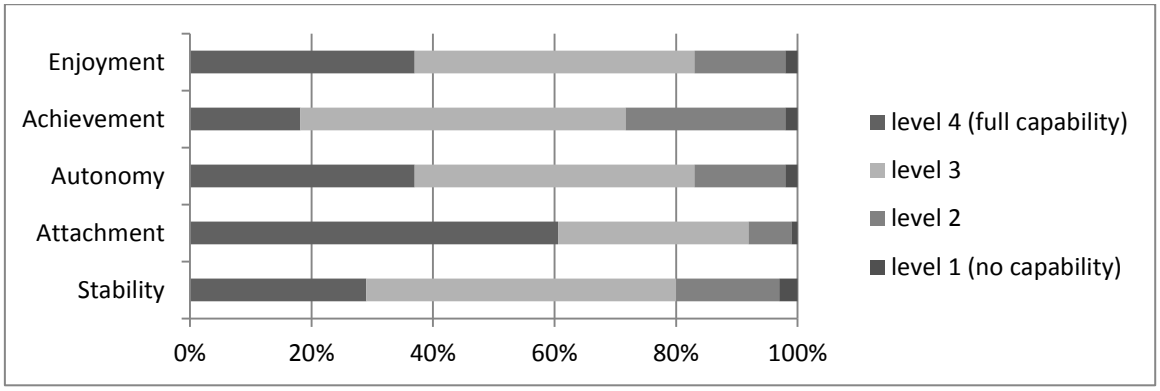
The study by Al-Janabi et al [116] reported a high completion rate of 99%. The study was administered face to face with participants who had agreed to be interviewed.

The feasibility of use of the ICECAP-A measure has also been assessed through qualitative think aloud methodology [181]. The study showed that informants understood and responded to questions in the ICECAP-A with a level of struggle and error comparable to the EQ-5D. Evidence also indicated that most people understood the questions were asking about their capability, rather than their functioning. This research provides positive evidence for the content and face validity of the ICECAP-A measure.

2.5.5.2. Response distribution

The response profiles of the ICECAP-A measure reported by Al-Janabi et al show notable difference between items in the percentage of respondents reporting full capability. Over 60% of people reported full capability on the Attachment item, while less that 20% reported full capability on Achievement (Figure 8).

Figure 8: Response distribution of ICECAP-A from Al-Janabi (2012)



2.5.5.3. ICECAP-A measures and socio-demography

No associations between sex or age and the ICECAP-A items were found by Al-Janabi et al [116]. Income and employment status was found to be associated with all ICECAP-A items, as was home ownership status. Financial worries were found to associate with Stability, Achievement and Enjoyment (see Table 4).

2.5.5.3.1. Relationship status and bereavement

Relationship status and recent breakup was associated with all ICECAP-A items other than Autonomy. Unexpectedly, however, recent bereavement was not associated with any ICECAP-A item.

2.5.5.4. ICECAP-A measure and health.

2.5.5.4.1. EQ-5D-3L

The EQ-5D-3L index score was found to be associated with all ICECAP-A items apart from Attachment. All EQ-5D-3L items were associated with all ICECAP-A items, with the exception of Mobility and Attachment. A number of these associations were not expected *a*

priori, and it should therefore be concluded that a closer association between the EQ-5D-3L and ICECAP-A than expected was found.

2.5.5.4.2. Other health variables

Having a long standing illness or being in receipt of informal care was found to associate with lower scores on all items apart from Attachment. Having an outpatient appointment in the last year was found to associate with Achievement and Enjoyment. No associations were found for inpatient admissions.

2.5.5.5. ICECAP-A and freedom

Al-Janabi et al [116] assessed the association between the ICECAP-A and three one-question assessments of freedom. As hypothesised, the ICECAP-A tariff score was positively correlated with these assessments, and these correlations were stronger than the correlations between the freedom variables and the EQ-5D-3L.

Table 4: Summary of primary hypotheses, results and conclusions from research articles referring to the ICECAP-A included in review

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures.
Al-Janabi (2012)	*Health* Impairments for physical health were expected to reduce capability for Stability, Autonomy, Achievement and Enjoyment, while impairments to psychological health were expected to reduce capability for Attachment. *Socio-demographic* Participants who were employed or had a good income, were expected to have higher Stability, Achievement and Enjoyment. Those in relationships were anticipated to have higher levels of Stability, Attachment and Enjoyment, but lower levels of Autonomy. Individuals with higher levels of education would also have higher levels of Autonomy and Achievement. Participants who had suffered a recent major negative life event (bereavement, break-up, financial problems, ill-health) were anticipated to report lower levels of Stability, Attachment, Achievement and Enjoyment. *Freedom* It was anticipated that correlations between freedom variables (opportunities, Control over what happens, can do things I want) and the ICECAP-A would be stronger than correlations	*Health* EQ-5D-3L physical health attributes were associated with capability on the Stability, Autonomy, Achievement and Enjoyment items. The anxiety/depression question on the EQ-5D-3L was associated with Attachment. *Socio-demographic* Employment status and income was associated with all ICECAP-A items. Relationship status was associated with all items apart from Autonomy. Education was associated with all items apart from Attachment. Associations were found between individuals who had suffered a recent break-up and all items apart from Autonomy, while financial worries were associated with Stability, Achievement and Enjoyment. Happiness was associated with all ICECAP-A items. *Freedom* Freedom variable showed stronger correlations with the ICECAP-A score than the EQ-5D-3L score.	The findings, which indicate that ICECAP-A capability scores are associated with freedom, socio-demographic and health variables, provide encouraging evidence of the validity of the measure in the general population. Evidence of some anticipated associations not being seen and some unanticipated associations being found tempered the positive conclusions of the authors.

	Author hypotheses	Results	A summary of author conclusions referring to the validity of the ICECAP measures.
	between freedom variables and the EQ-5D-3L.		
Al-Janabi (2013)	Qualitative work – no hypotheses provided.	A low number of "errors" (comprehension, retrieval, judgement, response) or "struggle" was seen. A number of informants demonstrated that the questions were asking what they could do, rather than what they did do. Some confusion was demonstrated over the use of the word "can" in the Attachment item.	With some degree of struggle and error individuals can respond to and answer questions about their capability. Distinctions can be made between their capability and functioning.

2.5.6. Discussion of methodological review findings

This review provides an initial indication that the ICECAP-O measure is feasible for use in patient and general population samples and that it measures what it purports to measure. The limited number of studies which provide evidence on the validity and feasibility of the ICECAP-A, means that there is less evidence on which to form a firm conclusion.

The ICECAP measures were designed to capture a broader conceptualisation of well-being, which frequently used health functioning or health-related quality of life measures do not capture. The assertion that the ICECAP-O offers a broader evaluative space appears to be supported by the initial research comparing the measure to health status and health-related quality of life measures. Studies identified by this review consistently reported moderate correlations with measures of physical health and disability. Furthermore, it was evident that some items of the ICECAP-O measure may not be associated at all with health measures. This suggests that while health is a factor that explains ICECAP-O scores, it is not the only or, possibly, even the predominant factor.

This tentative conclusion is supported by the results from studies which compared the ICECAP-O with socio-demographic variables and measure of general well-being. The results from socio-demographic variables showed that relationships and living with a partner had an influence upon ICECAP-O results. Limited evidence was also found which indicated that whether a person provided or received informal care had an influence on the ICECAP-O scores. These results, in conjunction with results that show measures of general well-being, life satisfaction and happiness to be strongly correlated with ICECAP-O scores, provide additional evidence that the measure offers a broader evaluative space.

Two sets of evidence suggest that the measure is feasible for use. The majority of response rates, in both patient and general population samples were in the high 90th percentile. Response rate is an important practical consideration for researchers considering using a measure. Secondly, the response profiles provided by a number of studies indicate that, with the possible exception of Attachment, the ICECAP-O measure items do not suffer from ceiling or floor effects, where responses cluster at one extreme of the score range. The Attachment item had the highest percentage of responses in the top level at 60%. To draw a firm conclusion about the absence of a ceiling effect a comparison between two measures of the same construct would be needed, but this was not possible.

2.5.6.1. Strengths and weaknesses of the review

This review used systematic search criteria to identify research pertaining to the psychometric properties of the ICECAP-A and ICECAP-O measures. Part of this strategy was the direct contact of a large number of researchers who have registered to use the ICECAP-A and ICECAP-O measures. This, in conjunction with a keyword/MeSH term search and forward/backward citation search, provides a high degree of certainty that the overwhelming majority of existing research was identified. While a comprehensive search was completed, and not a weakness of this review *per se*, the size of the existing pool of research on this topic is small. This results in some uncertainty and the need to draw tentative conclusions.

Due to the variety of outcome measures used, and differences in the reporting of results in the studies, a narrative reporting of findings was the most appropriate way to present results. It has allowed this review to draw out differences and highlight weaknesses in the research. As

the evidence base grows and a more uniform set of outcomes are reported, it may be possible to apply quantitative assessments to review findings.

The focus of this review was to identify and synthesise all existing evidence on the psychometric properties of the measure. This review has not made any attempt to assess the quality of research, or exclude research based on quality. Therefore, it is likely that studies of differing qualities are included in this review. When the evidence base is larger, future reviews may choose to apply quality criteria prior to synthesising the evidence.

2.5.6.2. Gaps in the literature

Research into the validity of the ICECAP measures is at an early stage. A smaller amount of research exists for the ICECAP-A measure than for the ICECAP-O, which is presumably due to the recent development of this measure. A distinction should be made between areas where the current evidence base does not allow firm conclusions to be drawn and areas where a clear gap (absence of research) exists. In general, further research should provide greater certainty for the tentative conclusions drawn in this review. Specifically, further research should shed greater light on the inconsistent results as to the effect of age (and other socio-demographic variables) and psychological health upon ICECAP scores. It should also further the evidence around the link between happiness and well-being and ICECAP scores.

Three primary gaps in the literature exist. First, all the studies identified by this review are cross-sectional studies, using data from one time point. As discussed in Chapter 3, for a measure to be considered valid, it must also be valid over time. Therefore, assessment of the sensitivity to change or responsiveness of these measures is an important future area for research. Second, while the ICECAP-O measure has been used with patient populations, to

assess differences between care pathways and differences between patients with differing severity of illness, neither of the measures have been tested in the setting of interventional research, specifically a randomised controlled trial. A measure, which is considered suitable for economic evaluation, must be found to be valid in such a setting. Third, the amount of qualitative literature on the content validity and feasibility for use of the measure is limited. Research using researchers and experts in the field would allow triangulation with the existing research using the general public.

2.6 Chapter conclusion

This chapter has identified the traditional methods for measuring quality of life in health research and randomised controlled trials. Efforts to put the capability approach into practice have been summarised, with particular focus on the ICECAP measures. The development of the ICECAP measures has been reported and the current literature pertaining to the validity of these measures has been reviewed. Three primary gaps in the literature have been identified. The research reported in Chapters 4 and 5 is designed to address the lack of qualitative research, while Chapters 6, 7 and 8 of this thesis were designed to provide evidence from longitudinal research and research within an interventional setting.

Before continuing to the empirical work, Chapter 3 reports the methodology by which validity and responsiveness should be assessed. The field of psychometric methodology is complex and has been under constant change since early assessments of validity. Debate still exists as to how best to assess the validity and responsiveness of measures. A summary of this debate is presented and the methodological norms which have been established in this area are reported.

CHAPTER 3. THE THEORY OF PSYCHOMETRIC TESTING

3.1. Introduction

This chapter will describe and discuss the psychometric properties of a measure: reliability, validity and responsiveness. The theoretical issues currently under debate in the research field of psychometric testing, and the practical considerations when assessing each property, will be covered. The properties of validity, and to a lesser extent responsiveness, have undergone considerable interpretation and re-interpretation, and are still in flux. A brief history of the evolution of theory is presented alongside current debates.

A discussion of the methodology normally used to assess reliability, construct validity, content validity and responsiveness is provided. This methodology is taken from leading works in the fields of quality of life research [138], measurement scale development [184] and psychometric measurement [185]. These methods are used in the majority of psychometric studies published in leading journals such as *Quality of Life Research, Value in Health and Health and Quality of Life Outcomes*. It is therefore representative of best and standard practice in the field and is the methodology which is used in this thesis to assess the psychometric properties of the ICECAP measures. The chapter concludes by looking at the challenges of assessing the validity of a capability measure in a health research setting; an area where there is little precedent in the literature.

3.2. Measurement

'The act of measurement is an essential component of scientific research'$_{(p.1)}$ [186]. Its importance has been long accepted in the health and medical sciences and more recently measurement has been seen as an important aspect of social science disciplines, including quality of life research.

When measuring quality of life, and other non-tangible concepts in social science, a focus is needed on the relationship between underlying (unobservable) concepts and observable indicators. Therefore, definitions of measurement such as 'the assignment of numbers to objects or events according to rules'$_{(p.22)}$[187], which may be appropriate in the physical sciences, are incomplete for use in social science. Here measurement involves both empirical considerations, the observable response or behaviours, and theoretical considerations, the underlying unobservable concepts [188]. Therefore an appropriate definition of measurement for work with quality of life is [188]:

> *'the process of linking abstract concepts to empirical indicants'*$_{(p.10)}$

The use of short questionnaires, which elicit the patients' perspective, can allow rigorous, reliable and valid measurement of quality of life and health-related quality of life. They are 'especially significant when symptoms, functioning, and well-being are important outcomes or areas of concern'$_{(p.s94)}$[189]. To have confidence in the measurement a patient reported outcome measure provides, information is needed on the reliability, validity, and responsiveness of the measure.

3.3. Reliability

Inconsistency is present in all observations and measurements [190]. Subtle variations in the measure, the manifestation of the construct of interest and the individual all contribute towards inconsistency. This inconsistency, or *error*, reduces confidence in a measurement and subsequently its usefulness. In order to have confidence in a measurement we must assess the degree to which a measure is compromised by error.

3.3.1. Classical test theory

The concept of reliability is closely linked with classical test theory. Classical test theory states that any measure or observation has two components: a true score, obtained in an error free observation, and an error associated with the observation [186,188,191]:

$$X = t + e \quad (1)$$

The term 'true score' is a misleading one [186] as true scores are hypothetical, unobservable values [188]. A person's true score is the mean score that would be obtained if the measure was given an infinite number of times [186,188].

All observations are affected by random error. Equation 1 states that the observed score will not equal true score due to error. Random error means that a score or observation will be higher or lower than the true score. Importantly, a score is as likely to be higher as it is to be lower (if a score is more likely to be either higher or lower, then this is systematic error and can affect the validity of the measure).

Classical test theory makes a series of assumptions about true scores, random error and the relationship between them: '1) the expected mean error score is zero; 2) the correlation

between true and error scores is zero; 3) the correlation between the error score on the measurement and the true score on a second is zero; 4) the correlation between errors on distinct measurements is zero'$_{(p.30)}$[184,191]. This can be summated in equation 2, where E represents the longitudinal mean [191]:

$$(2) \quad E(X) = E(t)$$

Equation 2 can then be written so that it refers not to a single observed score, true score and random error, but to the variance of these [191]:

$$(3) \quad VAR(X) = VAR(t) + VAR(e)$$

Based on these assumptions of the behaviour of true scores, random error and observed scores the reliability coefficient, which expresses the proportion of total variance which is due to 'true' differences between subjects is calculated in equation 4 [191]:

$$(4) \quad \text{Reliability} = \frac{\text{Subject Variability}}{\text{Subject Variability} + \text{Measurement Error}}$$

Equation 4 shows that subject variability will always be less than subject variability + measurement error. Therefore the reliability will vary between 0, no reliability and 1.0, complete reliability. Reliability is the proportion of true variance to observed variance; the greater the proportion, the greater the reliability.

3.3.2. Definition and description of reliability

Reliability is the extent to which a measure produces repeatable measurements, with a low component of inconsistency, or error [138,184]. This can be over time, between different modes of administration and between differing situations in which it is used. A measure

should yield consistent results within a population experiencing no change when assessed at different points in time and when assessed through postal questionnaire, online survey or face-to-face interviews.

'Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself'$_{(p.78)}$[192]. Reliability is population specific as it is an interaction between the situation, the measure and the population of subjects [186]. A test or measure cannot be reliable *per se,* as its reliability will vary between different populations of patients. Therefore when reporting reliability one should focus on test scores and report the reliability of the test with that population [186,193]. This thesis will use a slightly extended definition of reliability from Brazier et al [194]:

> *'Reliability is the ability of a measure to reproduce the same value on two separate administrations when there has been no change'$_{(p.66)}$ in the construct of interest*

3.3.2.1. Assessing reliability

There are two primary methods through which reliability can be assessed.

3.3.2.1.1. Test-retest method

The test-retest method, by which a measure is given at two different time points or through two different modes of administration, is an intuitively appealing way of assessing reliability [191]. Patients or participants whose condition is stable should be chosen and a time gap, which is neither too long nor too short, should be used [138]. If the scores are identical between test and retest the correlation will be 1, however the instability of measures will normally mean correlations across time points are less than perfect [191].

The retest method, which is expensive both financially and in terms of research time, has limitations. First, measuring an attribute of interest can induce change in that attribute [191]. After being asked at the first administration of a measure people can become more aware and be prompted to think about the domain on which they were questioned. Second, a "real" shift in the underlying theoretical construct may occur [184]. This can be controlled, to some extent, by selecting groups that have not changed (e.g. assessing participant for change using some other criterion and using a group that has not changed).

While these are ways that the test-retest method can underestimate reliability, a 'more common problem is overestimation due to memory'$_{(p.39)}$[188,191]. When the time interval between test and retest is short then subjects will remember their initial responses [189]. It appears logical that the shorter and simpler the test, the longer will be required before memory is no longer a factor. Nunnally [195] and others [196] purport that the current best practice of repeating the measurement within two-weeks to one-month is likely to bring in a strong influence of memory. It is suggested herein that brief patient-reported outcome measures should be tested for reliability using a longer re-test period, as the likelihood of memory of response is greater.

3.3.2.1.2. Internal consistency measures

Internal consistency is a measure of how interrelated the items of a measure are [138]. It is considered a measure of reliability, but its output provides very different information to that of a test-retest assessment: it provides information on how homogenous the items of a measure are. The method has the advantage that it does not require repeated tests such as the test-retest method [188]. Of this set of methods Cronbach's alpha [197] is a frequently used

option. Under this method inter-item correlations of a measure are recorded and a mean inter-item correlation is calculated. Using this calculation the Cronbach's alpha can be calculated using the equation 5 [191]:

$$(5) \qquad \alpha = N\bar{p} / [1 + \bar{p}(N-1)]$$

Where $\bar{p}$ equals the mean inter-item correlation and the number of items is represented by N [191]. Cronbach's alpha is therefore dependent on the mean inter-item correlation and the number of items in the measure; as both of these increase, so does the value of alpha. A weakness of the test, which needs to be considered when using Cronbach's Alpha, is that the output is dependent on the inter-item correlation, and also the number of items [138]: if either increase so will the score. The developer of the test recognised the need to control for this in commenting that 'a quart of homogenised milk is no more homogenised than a pint of milk'$_{(p.86)}$[138].

3.4. Validity

Validity is a fundamental consideration in the development and use of a measure [185]. This section will define, describe and discuss validity and the process of assessing validity, validation. First, it will start with a brief history of validity theory.

3.4.1. A short history of validity theory

The concept of validity, what we mean by it and how we measure it, has evolved markedly over the last century. At the centre of this evolution is a change from validating the measure towards validating the inferences drawn from the results of a measure [198]. This has not been a smooth evolution [199], rather periods of gradual change punctuated, and often reversed, by "landmark" texts that altered the accepted definition and assessment of validity [186].

Prior to 1954 validity essentially assessed how well a test estimated or predicted the variable of interest [200]. A test was considered valid if it measured what it claimed to measure [198]. The variable of interest 'was assumed to have a definite value'$_{(p.319)}$[200] measured by a criterion measure. Validity was measured through the correlation between the test and the criterion. This thinking can be seen in early validity texts [201] where a test was considered a valid measure of anything with which it correlated.

The lack of suitable *criterion* measures and the 'infinite regress' [200] involved in comparison to the "current" criterion (discussed later) meant that this early conception of validity assessment did not provide an adequate analysis structure. In his article entitled 'Stamp

Collecting versus Science' Frank Landy [202] described the pre-1954 period of validity testing as chaos consisting of 'dust-bowl empiricism'(p.1183).

In a landmark text Cronbach and Meehl [203] presented the findings of the 1954 American Psychological Association Committee on Psychological Tests. In what was later termed the "trinitarian point of view" [204], Cronbach and Meehl [203] presented a tripartite model for validity assessment. Alongside *criterion validity* the authors proposed *content validity,* an expert led examination of the content of the items of a test, and *construct validity,* a framework of hypothesis testing [203]. This "watershed event" [200] resulted in a departure from the test-criterion correlation assessment of validity and towards a scientific process of test validation [202] in four ways.

First, Cronbach and Meehl [203] advanced validity research by proposing the use of a nomological network (discussed later): a scientific theory which provides a framework for testing validity [198]. Using the nomological network the validity of the test is established through evidence that confirms the theory proposed. Scientific testing became integral to validity research and a scientifically sound, hypothesis based assessment of validity became the norm.

Second, validity assessment became more concerned with the inferences drawn from a test, rather than the test itself. Validity of a test or measure became a question of the accuracy of, and the degree of confidence that can be placed in, the inferences, or conclusions, drawn from the measure [184].

Third, Kane [200,205] proposed an argument-based approach to validation. It is now widely recognised that a test can never be validated absolutely, rather a reasonable case can be made

as to what the test scores mean [206]. This has resulted in validation becoming an extended analysis and an ongoing process, which includes the development process of the test and subsequent studies.

Finally, while Cronbach and Meehl [203] stated that the three forms of validity were categories within an overarching concept of validity, the notion of construct validity began to be seen as a general approach [200,207]. A unified version of validity, based around the construct validation methodology, was widely accepted [208].

These four points outline a modern, scientific, hypothesis based conception of validity which will be used in this thesis [200]. However, this concept of validation is not without its critics, with a further shift in the last ten years focusing on concern with the use of a nomological network [209–211]. The primary practical critique is that nomological networks of the sort required for construct validation do not exist [209]. The theoretical concerns focus on the overly complicated framework for construct validation and rejection of the interpretation of the test scores as a basis for validation [212]. This critique has attempted to bring the validity evolution full circle, reverting back to the assumption that 'a test is valid if it measures what it purports to measure'$_{(P.1061)}$[209]. Despite this debate, however, there are few methods associated with these views, and the remainder of the thesis uses established conventions for validity assessment.

3.4.2. Definition and description of validity

3.4.2.1. Definition of Validity

This thesis draws on a number of works [185,200,206,213] to define validity as:

> *The extent to which a measure allows us to draw accurate conclusions about the presence and degree of the attribute of interest, in an individual, at a given time, in a predefined context.*

This definition is now broken down in order to give a fuller explanation.

First consider the phrase *"the extent to which"*. Validity is not a binary phenomenon; it is not either present or absent, rather it will be present to varying degrees [199,214,215]. Therefore, validity is more accurately described through continuous rather than dichotomous indices [214,216]. The extent to which a measure is considered valid will depend on the amount and quality of research that has been undertaken, and the findings of this research [205,217]. Validity should be seen as evolving [199], with each new piece of research and evidence increasing or reducing confidence in the validity of a measure. In the initial stages of validity assessment the most that can normally be said is that a measure shows indications of being valid to some degree.

Second, *"to draw accurate conclusions"* refers to the proposition that we do not validate a scale or measure, rather the interpretation of the results it yields [218]. By validating the inferences or conclusions that we can make based upon results, modern validation is also examining the theoretical basis of the measurement. Kane [212] highlights the importance of this through introducing the "begging-the-question fallacy"$_{(p.50)}$, where conclusions are drawn which go beyond the more modest conclusions which were validated. To illustrate, take the example of a school physical education test of whether a student can run one mile[2]. If this test is used to report whether a student can run one mile, then this conclusion could be accepted at face value with relatively little validation. If the test was used as a measure of the construct "physical endurance" then a definition of physical endurance and an explanation of why this

[2]Example adapted from work by Kane [212].

test will show this, would be required. If this test was being used to draw conclusions about athletic ability and whether the student will compete at future Olympic games, then in the absence of considerable supporting evidence, this conclusion could not be considered valid. The interpretations or conclusions drawn from the test are important and it is these that need to be validated [212].

Third, the ability of a test to provide information *"about the presence and degree"* of an attribute is vital. There are few situations where a test that shows simply the presence of an attribute of interest would be useful. Even in conditions where the state is binary, such as pregnancy, further information is normally useful. Clinicians, scientists and social scientists generally require measures that show the degree to which a construct is present.

Fourth, it is important that the scale measures the *"attribute of interest"*. A scale may accurately measure an attribute which is not of interest [188]: a depression measure needs to accurately measure depression, rather than lack of happiness or lethargy. The content of the measure needs to be an adequate reflection of the construct to be measured [219]. Using the example of the one-mile run test above, in the first conclusion drawn the test is the whole of the construct, while in the second and third conclusions the test is to varying degrees a less adequate reflection of the construct being judged. This is important with respect to to the conclusions and inferences that can be drawn.

Lastly, a measure is only valid "*in an individual, at a given time, in a predefined context".* A measure that shows high validity within a select demographic, in a distinct circumstance, might not be as valid in a different circumstance or with a different demographic. This is why, for example, there are numerous publications validating the EQ-5D-3L measure in

different age patients, with different illnesses and in different countries with different cultures [220–224].

3.4.2.2. The process of validation

Validation is an ongoing process by which, incrementally, evidence is brought to bear on the validity of a measure [200]. This thesis will use a definition of validation from Lawshe [225]:

> *'Validation is a procedure, process, or strategy whereby we collect or generate data to determine or defend the extent, degree or strength of the inferences that can be made from a set of test scores'.(p.237)*

This definition recognises validation as a scientific process of extended, hypothesis-based investigation through which a sound validity argument is formed to support the interpretations of the test scores. Hypotheses should be stated, along with the reasoning or evidence behind them. They should be subjected to testing and the result should be reported. Whenever possible different types of validation should be referred to and not validity types [225,226].

The argument-based approach to validation proposed by Kane [205] requires the generation and collection of evidence to test the proposed interpretations of the results of the measure. Kane [205] noted that validation was a matter of building a validity portfolio, based on the proposed use of the test, which could be used to justify the interpretations drawn [205]. Ambitious interpretations generally require more evidence than narrow interpretations [212]. Whatever the interpretation, compilation of the validity portfolio normally starts during the development of the measure, with specification of the expected uses of the measure. Once developed, a critical arm's length analysis should occur, with assumptions and inferences subjected to testing.

Cronbach and Meehl [203] stated that one finding different to expectations would be enough to conclude a measure is not valid. However, it is considered here that negative evidence should be handled in the same way as positive evidence, as incrementally informing the investigation. There are three possible reasons for a negative result [227]: First, the test does not accurately measure the attribute of interest. Second, the theory which generated the hypothesis for testing is not correct or incomplete. Third, the experimental design of the validation study failed to test the hypothesis properly, which can include the criterion being faulty or inappropriate (discussed later). It should also be noted that any combination of these may also occur. Understanding the reason for a negative result is important when planning follow-up studies.

3.4.3. Criterion validation

Criterion validation is an analysis of the degree to which the scores of the new measure that is under consideration are a satisfactory reflection of an already accepted measure of the attribute of interest, ideally the "gold standard" used within the field of interest [186,219]. The degree of correspondence between the two measures is indicated by the strength of their correlation [188]; this is the only evidence that is relevant. *Criterion validation* is further split into two types: *concurrent validation* and *predictive validation* [227]. If the new measure and criterion measure are administered, and the scores determined, at the same time then this is *concurrent validation* [188,227]. Concurrent validation is normally used when examining a proposed replacement for an existing measure [227]. *Predictive validation* occurs when a future criterion is compared with the new measure [227].

There are a number of methodological issues with the use of *criterion validation.* First, if there is an existing criterion or "gold standard", why is another measure or test required? The reason may simply be that the new test is less costly or quicker to administer, or that the old test was in some way invasive and an alternative is needed. If the reason is that the developers of the new measure consider the old measure flawed to some degree, then use of the old measure as a criterion is then methodologically flawed [138].

Secondly, in many areas of study there is no adequate criterion to compare a new measure with [200,228]. This is often the case when seeking to validate patient-reported outcome measures, or new measures such as capability measure, in health [117,189]. When no criterion measure is available criterion validation is clearly not a viable option [191].

Third, correlation evidence is the only applicable evidence for criterion validation [191]. This can lead to some spurious and atheoretical conclusions. For example, if car ownership correlated strongly with university dropout, then this would be the 'whole story' from a criterion validation aspect [191]. A mistaken assumption of criterion validation is that correlation equals causality [209].

Fourth, criterion validation is highly dependent on the availability, quality and validity of the criterion itself [200]. If the criterion is a weak approximation of the underlying attribute of interest then a validation study using the criterion has to be judged as unsound. As Cronbach [218] stated 'all validation reports carry the warning clause, 'insofar as the criterion is truly representative of the outcome we wish to maximise''$_{(p.488)}$.

Finally, through the use of criterion validation, there is the problem of an "infinite regress" [226]. Take, as an hypothetical example, the measurement of bodily health. The original

measure has a "true" correlation with global, bodily health of 0.9. Three generations of measurement development occur. Each new measure records a strong correlation of 0.9 with the previous. In doing so it becomes the "gold standard" and hence the test which is used as the criterion in future analyses. After three generations of test development we could be left with a measure which has a "true" 0.66 correlation with global health (.9*.9*.9*.9 = 0.6561).

The weakness of criterion validation has resulted in increasingly limited use. In an investigation by the International Society for Quality of Life Research into minimum standards for patient-reported outcome measures only 10 percent of informed participants stated that it was important to have criterion validation evidence before using a measure [229]. Borsboom [209] has described it as an 'atheoretical, empiricist idea…[which]…was one of the most serious mistakes ever made in the theory of psychological measurement'$_{(p.1065)}$. If used at all, criterion validation should be viewed as contributing incomplete and potentially compromised evidence to the validity argument.

3.4.4. Construct validation

Construct validation was proposed by the APA Committee on Psychological Tests in 1954 as an alternative to criterion validity. Cronbach and Meehl [227] suggested that it should be used in cases where the attribute of interest is not "operationally defined" recognising that when a non-tangible concept is being measured, the validation of the measurement instrument is difficult. Attributes such as height and weight are readily observable and easily measured. Other, non-tangible attributes such as depression, happiness, quality of life and capability are not easily observable, and therefore not easily measured.

In these circumstances Cronbach and Meehl [203] recommend that a nomological network should be used. A nomological network is an interlocking system of laws and relationships used to define the links or associations between constructs [217]. The attribute being measured is placed in this network as a construct. 'A construct can be thought of as a mini-theory to explain the relationships among various behaviours'$_{(p.257)}$[186]. The proposed system of laws and relationships within the nomological network is used to generate specific, testable hypotheses [203]. A measurement can then be validated based on the laws detailed in this nomological network. For example, if within a nomological network for happiness wealth is not a determinant of happiness, the scores from a measurement of happiness should not systematically differ between people of different wealth. Within a well-designed study this can be tested. The aim of construct validation is to embed a measurement of a construct within a nomological network and test the relations with other variable within that network [230]. In a practical sense construct validation is [231]:

> *'a series of procedures concerned with assessing the extent to which the dimension scores of an instrument correlate with other hypothesised measures or indicators of the health concept or concepts of interest'.*$_{(p.43)}$

3.4.4.1. Construct validity as a unifying force

Construct validation has come to be seen as a general organising concept for the assessment of validity [226]. Four central principles of construct validation have emerged as general principles of "good practice".

First, the construct validation model requires an extended analysis, involving theory development, development of evidence-based hypotheses, development of measurement procedures and continued testing using the formed hypotheses and procedures [226,227]. Inherent within these requirements is a criticism of the shorter forms of content and criterion

validation methods, which often rely solely on expert opinion or a single validity coefficient. Consequently criterion validation is now little used and recent work on content validation, which is presented below, is beginning to emphasise the need for the triangulation of data and work with different informants [232,233].

Second, there is a need within construct validation to clearly define the theory and the construct under examination. As Cronbach and Meehl stated 'a necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least *some* of whose laws involve observables'$_{(p.290)}$[203]. Kane has provided forceful justification of this requirement by stating 'validation is difficult at the best, but it is essentially impossible if the interpretation to be validated is unclear'$_{(p\ 329)}$[226].

Third, validation should be viewed in as the process of constructing a validity "argument" [205,226,227] through a continuous process [207,234] of building a validity portfolio for a measure [185]. Finally, it is the interpretations of the test scores which are validated, not the tests themselves. Validity is a property of the interpretations and uses of the test, not of the test per se [212].

'Taking construct validity as the unifying principle for validity puts validation squarely in the long scientific tradition of stating a proposed interpretation clearly and subjecting it to empirical and conceptual challenge'$_{(p.325)}$ [226]. This has broken down the compartmentalisation of different validity types [207], which caused confusion and resulted in the opportunistic use of the different types of validity [235]. Different types of validation efforts are needed for different types of interpretative arguments [212,236] and the distinction between content, criterion and construct continues to provide a structure through which this

can happen. However, due to the idea of construct validation these now happen under a unified banner of "validity" [200].

3.4.4.2. Assessing construct validity

3.4.4.2.1. A five point model for validation

Drawing heavily on proposals by Kane [200] and Smith [234] and based on the acceptance that construct validity provides a model of best practice in validation that can be applied to all validation types, a five point model for validation is defined.

1) Specify the theoretical background as clearly as is allowed by past empirical research and theoretical work. The theoretical construct in question should be carefully specified.
2) Develop and describe the validity argument by stating hypotheses to be tested. Evidence based rationale should be presented where possible.
3) Specify the research design to be used to test the hypotheses. Where useful validation "types" should be referred to.
4) Examine the stated hypotheses by subjecting them to testing. If there are parts of the validity argument which are considered more problematic, a focus should be maintained on these.
5) Interpret the empirical evidence.

Steps 2, 3 and 4 can be cyclical, with the interpretation of the empirical evidence informing a continued development and description of the validity argument.

Three methods dominate construct validation efforts: known groups analysis, convergent analysis and divergent analysis. *Known groups validation* is a simple form of validation which should be completed alongside other forms of validation [138,184]. This analysis can assess whether a measure is capable of distinguishing between two (or more) distinctly different groups, but not of distinguishing between people with subtly varying degrees of an attribute [138]. Under this process two groups are selected which are known to have differing amounts of the attribute of interest [231] or differing amounts of a theoretically related attribute. For example: when validating a measurement of happiness, known groups validation would look for the ability of a measure to discriminate between groups of people which are happy or sad. Or, if marital status was hypothesised to be a determinant of happiness, then scores on the happiness measures should be significantly different between the two groups. Validity is indicated by a measure being able to show differences in the expected direction and it is the magnitude of this difference, rather than the significance of this difference that is important [138].

A methodological issue exists around the selection of known groups. In delineating the groups a judgement has to be made. If an existing measurement will be used to select the groups, the validity of that measurement is an important consideration. If a new measure is being developed because of some perceived inadequacy with the existing measure, then in using the existing measure to select the groups it must be recognised that the selection of groups is likely to be in-adequate [186]. If there is no available measure, and the groups have to be intuitively defined, then confidence becomes weaker still.

Convergent and *divergent validation*s are other options in construct validation. Through convergent validation information on validity can be gained by seeing how closely the

measure is correlated with other variables of the same construct [186] or variables that are hypothesised to relate to the measure. The direction and strength of the correlation are important for understanding the association between the measures. For example, when validating a measure of happiness we might hypothesise that friendship is a determinant of happiness and the measure under consideration should therefore correlate with a measure or indicator of friendship. A positive correlation shows that that the more friendship someone has the higher the scores on our happiness measure. We might hypothesise this relationship to be a strong one, therefore strong correlations would provide confirmatory results. However, a very high or perfect correlation would suggest the measure of happiness is simply measuring friendship [138].

Divergent validation is required to ensure that a measure does not correlate with attributes that it is hypothesised to be independent from [138]. If happiness is hypothesised to be independent of wealth, then there should only be (at best) very weak correlations between our measure of happiness and wealth [186]. It is important to note that if an inverse relationship between a measure and attribute is hypothesised, this would be assessed under the banner of *convergent*, not *divergent, validation.* In this situation a correlation would be hypothesised, whereas for divergent validation little correlation between two measures would be hypothesised.

3.4.5. Content Validation

A great deal of variation can be seen in the way content validation is defined in the literature. A basic definition is 'the degree to which the content of an…instrument is an adequate reflection of the construct to be measured'$_{(p.743)}$[219]. This definition can be extended to focus

on ‘the extent to which a scale or questionnaire represents the most relevant and important aspects of a concept’(p.743)[233]. Furthermore, assessment of content validity can provide ‘evidence that the conceptual framework, content of items and overall measurement approach are consistent’(p.1263)[232].

These definitions are based on the premise that the items of a measure relate to an underlying concept [233]. Implicit is the assumption that a test or measure cannot sample the full domain of content that is relevant to a construct, but only specific areas that are centrally relevant. When assessing content validity the need to specify the full domain or theory is paramount [203]. The investigation then needs to focus on whether relevant aspects have been sampled, so that content validation furthers an understanding of the inferences which can be drawn from the results of the measure; if a measurement does not fully assess a concept then this needs to be understood so that inappropriate inferences are not drawn [184]. This places content validity within the rigorous approach to validation described above.

A subsection of content validation is face validation: the assessment of whether the dimensions within the measure are sensible and appropriate [231] and whether the measure appears to be valid [237]. While highly subjective, it is desirable for a measure to have a high degree of consumer acceptability [237] as selection for use is vital if the test is to prove useful to the research community. If a test does not appear to be valid it is entirely possible that clinicians will refuse to use it and, if the measure is used, subjects may pay it little attention as they deem it irrelevant.

3.4.5.1. Assessing content validity

Content validation and face validation differ notably from the other forms of validation as they are based largely on the judgements of individuals, whether they are patients, public or research professionals [184]. The focus of content validation is on the content of the measure, normally before scores are collected. Whereas quantitative methods are used to assess criterion and construct validation, 'the most appropriate way to collect data to support content validity is by conducting qualitative research'$_{(p.1263)}$[232].

A report from a working group meeting of the Patient Reported Outcomes Measurement Information System (PROMIS) initiative noted an inconsistency in the evaluation of content validity [233]. This, coupled with the perception amongst some that qualitative methods are "a soft science" [238], requires that rigorous methodology is employed to maintain the scientific integrity of this analysis [232]. There is currently a continued effort to improve, and where possible standardise, content validation [232,233]. The central points of this effort are focused around the methodology employed, the populations used and the triangulation of data.

The methodology employed to assess content validity should be meticulously documented and transparent. The research should be grounded in the data, rather than following assumptions made at the outset [239]. The analysis should be iterative, thematic and constantly comparative [240,241]. Analysis should occur throughout the data collection process and findings of the analysis should be used to inform and improve future interviews. Different methodology will be used depending on whether the focus of the work is to develop

a measure with valid content or to assess the content of an existing measure, however these central points should be maintained [232].

The population used for qualitative research is central to content validation. Informant selection should be achieved through purposive theoretical sampling [239,242,243], to achieve a maximum variety within the sample or to achieve a selection of individuals based on some characteristic. These methods can be used to select both patients or public[3] and expert informants. Patients can provide an insiders' – emic – perspective and have first-hand, personal experience of both the concept and how it might be affected by different situations [233]. For example: a cancer patient would be able to describe how the disease has affected their quality of life. Experts can provide an outsiders' – etic – perspective [233]. For example: an oncologist might have a broad knowledge about the different ways cancer has affected their patients. As with the methodology employed, the use of experts versus patients might vary depending on whether a measure is being developed or validated post-development. However, to reach a secure judgement on the validity of a measure both patients and expert opinions should be elicited [233].

The triangulation of data involves 'the use of different methods and sources to check the integrity of, or extend, inferences drawn from the data'$_{(p.46)}$[244]. This can improve the external validity of findings and strengthen confidence in the conclusions drawn [242]. Data and findings from one to one interviews and focus groups should be checked against each other, as should data from expert versus patient informants and when and where possible from qualitative versus quantitative research.

[3] Patients may be more appropriate for validating a capability measure in a health setting, while the general public may be more suitable for capability measures to be used in other contexts.

3.4.6. Feasibility

To draw accurate conclusions from the results of a measure, the measure must be practically useable in the relevant context. Feasibility can be assessed through both quantitative and qualitative methodology. Completion rates and missing values, time taken to complete a measure and participant and/or researcher post-completion ratings of difficulty, can all provide information on the feasibility of the measure [245]. Qualitative methodology can be used to elicit opinions of the measure from both participants and researchers. This can be done through semi-structured interviews or through the use of methodology such as "think aloud" [181]. The triangulation of quantitative and qualitative data can provide a fuller assessment of the feasibility of a measure than use of either method on its own.

3.5. Responsiveness

3.5.1. The measurement of change

In health, as in the fields of psychology and education, change is measured for three main reasons [186]. First, to identify the differences between individuals in the amount of change that has occurred. Clinical researchers may want to identify individuals who are responsive to a treatment or therapy. Second, once individual differences in change are measured, one may then be interested in the correlates of this change. The third reason, and the focus of most clinical trials, is to infer treatment effects from group differences. Under this goal the average change for a control group and intervention group is usually compared to determine the presence of a treatment effect.

3.5.2. Definition and Description

The terminology "sensitivity-to-change" and "responsiveness" are often used interchangeably [246]. However, the meanings of each differ importantly and describe different properties of a measure.

3.5.2.1. Sensitivity-to-change

Sensitivity-to-change is the ability of the measure to record a level of change in the construct being measured. The concept of sensitivity to change makes no judgement about whether the change is meaningful, either clinically or to the patient. This thesis will use a slightly extended version of Laing's [246] definition of sensitivity-to-change:

> *'The ability of an instrument to measure change in a state regardless of whether it is relevant or meaningful to the decision maker'* $_{(p.II\text{-}85)}$ *or patient.*

3.5.2.2. Responsiveness

Responsiveness refers to the ability of an instrument to measure important or meaningful change [186] and is a key psychometric property of a measure [194]. Responsiveness is therefore an interpretation of change that occurs in the score of a measure; a question of what is meaningful. This change may, for example, allow an individual to achieve an essential task of daily living, or live with a manageable level of pain [247]. Therefore, like validity, responsiveness is context specific and a matter of the interpretation of change in the score of a measure [246]. A description of the brief history of how change is interpreted and meaningful change defined is followed by the ways in which it is currently assessed.

3.5.3. A brief history of the measurement of change

In 1987 Guyatt et al proposed the minimal clinically important difference as a method through which change over time could be interpreted [248]. This was followed two years later by a landmark paper from Jaeschke et al [249] who developed the proposition of minimal clinically important difference by defining it as 'the smallest difference which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management'$_{(p.408)}$. In these early interpretations of change the patient was placed centrally in a judgement that was taken in a decision making context [250].

Guyatt et al [251] then proposed a change in terminology from minimal *clinically* import difference to minimal important difference, which the authors defined as 'the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient's management'$_{(p.172)}$[250]. The differences between the Jaeschke definition and Guyatt's are

small. The "C" for clinically has been dropped, but it is still implied within the definition, and importantly it specifies that the change must occur in the domain of interest and not unspecified general change [250].

A further extension to the definition was added in 2005 by Schunemann: 'the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and would lead the patient or clinician to consider a change in the patients' management'$_{(p.594)}$[252]. These authors stated the importance of patients being informed and that informed proxies could be used when required [252]. Therefore, current definitions of minimal important difference, through which change over time can be interpreted, normally refer to the following points: it is a measure of the minimal difference in the outcome; patients or proxies are the judge of difference; the definition is placed firmly in a decision making context; and while clinical is now removed from the title, clinical management is an important contextual consideration.

3.5.4. Responsiveness as longitudinal validity

Disagreement exists within the literature as to whether responsiveness, along with validity and reliability, is the third vital psychometric property [253] or whether it is a part, or a sub-section, of validity: longitudinal validity [186,215]. This is an important consideration. If responsiveness is taken to be a separate measurement property then a measure can be said to be valid, but not responsive. If responsiveness is taken to be part of validity, then a non-responsive measure will be considered to have a low level of validity. Simply put 'a measure valid at one time point should also be valid at another time point'$_{(p.74)}$ [215].

Responsiveness is also context specific as it assesses the minimally important difference to an individual or population in the context which they are in. Therefore, if we say that a measure validly measures a construct in the population, then we also have to assume that it is able to respond appropriately to minimally important change in the construct in an individual. 'To maintain otherwise is to claim that an initially valid instrument may somehow lose its validity over time'$_{(p.74)}$ [254].

This reasoning then states that evidence of the responsiveness provides information on the validity of the measure. Borsboom states that 'a test is valid for measuring an attribute if variation in the attribute causes variation in the test scores'$_{(p.1067)}$[209]. Understanding whether and how variations in the attribute produce changes in the measurement outcomes, increases our understanding of the validity of the measure. Hays and Hadorn [215] are correct in stating that responsiveness should be considered as part of the validity argument portfolio of a measure. This approach will be adopted in this thesis.

3.5.5. Floor and Ceiling Effects

A measure should be responsive throughout its whole range. Floor and ceiling effects result in less sensitivity at the extremes of the top or bottom end of the measure [255]. This will normally happen when a large proportion of respondents select the top or bottom level of a measure, or item within the measure. When this occurs the measure has lost the ability to detect improvement in a large section of the population (i.e. there is no level which to "improve to"). However it is often hard to quantify a ceiling effect.

An assessment of the percentage of respondents answering the top level of an item or reporting full scores on a measure is often used to indicate the presence of a ceiling effect.

McHorney and Tarlov [256] have suggested that if 15% or more of respondents answer the top level then a ceiling effect may be present. However, using a percentage of respondents in isolation does not provide a definitive indication of a ceiling effect. A fuller assessment of ceiling effect is to compare the results of one measure with results from a measure of the same construct [154].

For example, in many populations the majority of respondents will answer the top level of the EQ-5D-3L items [257,258], with (for example) over 95% of a population registered with a general practice saying they are not anxious or depressed. In this situation it is reasonable to presume that not all 95% have perfect psychological health; rather they that didn't feel that the next level down (moderately anxious or depressed) applied to them. This assumption can be assessed through the use of a comparator measure of anxiety and depression. Through getting the same respondents to answer another measure, those reporting no problems on the EQ-5D-3L anxiety and depression item can be assessed to see what they report on another measure. This comparator measure should, based on existing evidence, be considered to be more sensitive and not suffer from ceiling effects. If the majority of those scoring the top level on the EQ-5D-3L item also score highly on another measure, then this might indicate a small ceiling effect; the measure has appropriately categorised people as having no anxiety and depression. If a minority of the respondents scored highly on a comparator measure this indicates a large ceiling effect; those stating they had no anxiety or depression have been shown to have some degree of problems on another measure.

3.5.6. Assessing responsiveness

When assessing the responsiveness of a preference-weighted measure (e.g. EQ-5D) consideration needs to be given to the effect of preference weights themselves. Responsiveness analyses are assessments of the descriptive system of a measure as well as the overall outcome of the measure. This is important in preference- based measures as the ability of the descriptive system to detect change in a construct is an essential precursor for the ability of the measure to reflect preferences [259]. However, the use of preference scores in the responsiveness analysis may result in conclusions being drawn about the responsiveness of the descriptive system of the measure based on the preference values attached [259]. For example, a change of, say, 10% in the descriptive system may be considered a sizable change. However, it might be of little value once preferences are attached. If the analyses was just completed using the preference-weighted scores, the conclusion could be made that the measure is not responsive; when in fact the measure is responsive, but the change is not valued. To avoid this situation and to provide a full analysis of the responsiveness of the measure, preference weighted and non-preference weighted scores should be used [259].

3.5.6.1. Anchor-based assessments of change

The taxonomy of methods for assessing responsiveness can be summarised under the headings of anchor-based and distribution-based approaches [260]. *Anchor-based approaches* explore the relationship between the change in scores of a measure and the same or similar concept measured by an independent anchor [260]. When using the anchor based approach it is recommended that multiple anchors are used [261]. Ideally a combination of independent

clinical and patient related anchors across a number of different patient populations should be used to assess the responsiveness of a measure. Potential anchors fall into 3 main categories: patient ratings, clinician ratings or objective clinical measures.

A patient ratings method often uses a global transition question or a global rating of change question [250]. Here, patients are asked to rate their overall change in the construct under consideration between two time points. Patients answer on a gradient scale, which normally has between 5 and 15 points [260], as to whether they are "better", "worse" or "about the same". Informants are then categorised into groups which have improved, deteriorated or experienced no change. Groups that have changed a little, by the one point on the gradient scale, can be used to assess the minimally important difference. These categories indicate the threshold at which patients begin to notice differences as those in these groups have experienced minimal change [261]. Differences between the groups in the score on a measure are then assessed.

The primary concern with the use of global rating of change is the reliance on patient memory, which can be inaccurate and suffers from systematic error in the estimation of past and current health state [260]. This error is apparent when, as noted by Walters and Brazier in their paper [262], change rating has a strong correlation with the follow-up measurement and a very weak correlation with the baseline measurement. In response to this concern the FDA proposed the use of the patient global rating concept [260]. Under this method patients rate their current state on the construct at baseline and follow-up, changes in ratings are calculated across time points and groups of participants are formed based on this calculation. For example, groups may be those who have improved, those who have not changed and those who have worsened in the anchor measure. When using this method it is important to identify

those patients who have experienced minimal change [261]. Clinician ratings can be calculated in much the same way as a patient rating described above.

Objective clinical measures are an important tool for interpreting change in a measure as they link change in the construct with change in an external criterion [260,261]. Clinical measures need to be relevant to and correlated with the measure under consideration. For example: there would be little point in assessing the responsiveness of an asthma related quality of life questionnaire against a clinical measure of joint mobility, however this clinical measure might be very relevant for an arthritis related measure of quality of life, or a generic quality of life measure. The change in the measure under consideration can be interpreted against large, medium and small changes in the clinical criterion. As when using patient global rating of concept it is important to identify those patients who have experienced minimally important clinical change.

It is strongly recommended, and widely accepted, that multiple anchors are used when completing an anchor based responsiveness analysis [251,260,263]. Anchors should be selected that have a theoretical or proven association with the measure under investigation. Selection of anchors can be informed by using an initial assessment of the correlation of change scores to identify measures with an acceptable association with the measure under investigation. An acceptable correlation threshold is taken to be 0.3 [260,261]. Revicki [261] states that alternative (lower) correlation thresholds may be acceptable in some situations. Other factors taken into account can include: a) cross-sectional correlations at baseline and follow-up between the measures, b) theoretical or methodological reasons for using the anchor or c) whether analysis using the anchor would increase the understanding of how a

measure responds with changes in the anchor, which may be of importance to investigators and researchers.

It should be noted that Revicki's rule of thumb [261] could produce a cyclical analysis process that may increase estimates of responsiveness. When assessing the relationship between the changes of two measures, through an anchor based analysis, only using measures that have already been shown to correlate with each other may produce inflated estimates of responsiveness. This has the potential to manufacture estimates which are more positive than may otherwise be the case. Therefore, the consideration of additional factors detailed above is important.

3.5.6.2. Distribution-based assessments of change

The distribution-based approach is a set of methods for estimating change based on a statistical parameter of the population or sample [260]. Normally this statistical parameter is the relation between the magnitude of effect and some observed variation or distribution within the sample [261,264]. The primary criticism levelled at distribution based approaches is that they are "anchor free" and have no external reference point [250]. Some have concluded that this makes them "meaning free" [261] and so provides no direct information about the MID. Given this limitation, FDA guidance for patient reported outcome measures suggests that distribution based approaches should play a supporting role [260]. There are two frequently used distribution based approaches: the the effect size and standard response mean, which are discussed at greater length in section 6.4.4.4. Effect sizes and standardised response means.

3.6. Challenges of assessing the psychometric properties of the capability measures in a randomised controlled trial

Discussed above are the three psychometric properties and methods for assessing these properties. The chapter is based on key texts and literature from the fields of psychology, health outcomes or patient reported outcome measures [138,184]. Little, if any literature exists on the methodology of validating capability measures. However, there are a number of challenges when validating measures of capability in the health research setting or a randomised controlled trial.

3.6.1. Reliability

To assess reliability via the test-retest method a stable population and two time points close in time at which the measure can be completed are required. Assessing reliability (of any measure) in a health intervention setting, such as a randomised controlled trial, is therefore methodologically challenging. Where those in the population may have some form of health condition and the presence of an intervention is looking to bring about change, the population is not likely to be stable. A solution to this problem may be to use measures included in the research to identify a stable sub-section of the population. For example, when assessing the reliability of a health measure, using other measures of health to identify a population which has not changed may offer a possible solution. However, as discussed below, indicators of capability are not normally included with trial questionnaire packs so there is less scope for this solution when dealing with capabilities.

The 2-4 week gaps in trial follow-ups required for assessment of reliability are rare within an intervention setting where the aim is to look for change rather than stability.

Internal consistency, which is often considered a form of reliability, makes the assumption that a measure contains a number of items that assess a single, clearly delineated construct (e.g. physical health or happiness). As discussed in chapter 1, a strength of the capability approach is the number and breadth of functionings that are held to be relevant. A capability measure may assess a number of different constructs under the heading of capability. Therefore a relatively heterogeneous sample of questions may form a capability measure. If this is the case, a measurement of internal consistency, such as Cronbach's Alpha may be inappropriate or have little informational value.

3.6.2. Validity

There are two main challenges when seeking to validate a capability measure in a health research setting. First, there is the need for a clearly defined nomological network, in the form of hypotheses which can be subjected to testing. A number of hypotheses need to be formed about the ICECAP measure's association with indicators of health. As discussed above, capability is an under-defined theory, which has only relatively recently been applied to health research. In comparison to a construct such as health-related quality of life there is little developed theory and only embryonic evidence of how capability may relate to health constructs. Applying construct validity when there is not a well-defined, solid theory is challenging [200]. Kane [226] and others warn about the dangers of "weak programme" validation, which is construct validation without the specification of hypotheses.

A second challenge is that the majority of questionnaires administered in baseline and follow-up questionnaire packs are measures of health, disability or pain. These questions allow easy construct validation of a measure of health or health-related quality of life, but present a challenge when validating measures of capability. A health-related quality of life measure would be expected to have a strong association with other measures of health; however a capability measure may be expected to have a weaker association. Some aspects of capability may not be expected to associate at all with measures of health. For example, being in pain may have very little bearing on whether or not someone can have friendship or feelings of agency and worth. This consideration should be reflected in the hypotheses drawn. Furthermore, the type of measures in a trial questionnaire pack are unlikely to allow a comprehensive validation of a capability measure. Health is only one of a number of factors impacting upon a person's capability.

3.6.3. Responsiveness

A number of methodological challenges exist when seeking to assess the responsiveness of a capability measure in a trial setting.

The majority of responsiveness analyses assess how the scores of a measurement instrument change when the construct, that the instrument is designed to measure, changes. In line with this aim anchor-based methods normally use an anchor measuring the same, or a very closely related construct. The absence of alternative measures of capability makes the formation of an anchor challenging, and in health research in practice is likely to be limited to anchors related to health.

This limitation has some important implications for the analysis. Two requirements have to be met when choosing anchors: first, if the anchor is not a measurement of capability it needs to have a theoretical relationship with the capability. This proposed relationship needs to be clearly described. Second, the anchor should have appreciable association with the measure under consideration [260]. Therefore, when selecting anchors associations and correlations at baseline, and over time, need to be taken into consideration.

The expectations about how much a capability score will change in response to a change in health need to be managed. Large changes in capability with changes in health and strong correlations between capability measures and health measures are unlikely. A number of different factors affect the capability of a person, health being one, and therefore the effect of changes in health upon capability will be moderated. Furthermore, the absence of capability measures and dominance of health focused measures in health research will impact upon the conclusions that can be drawn from a responsiveness study in such a setting. Firm conclusions may be formed as to how a capability measure responds to changes in health, but there will be less certainty in the conclusions as to how the measure changes with changes in capability.

3.7. Chapter summary

This chapter defines the psychometric properties of reliability, validity and responsiveness and summarises the methods through which they can be assessed. The challenges of assessing the psychometric properties of a capability measure in health are discussed. Chapters 4 and 6 is a discussion of the steps taken to address the challenges identified. Chapter 5 reports the outcome of the qualitative assessment of the content validity of the

ICECAP-A measure, while Chapters 7 and 8 report the outcomes of the quantitative assessment of the validity and responsiveness of the ICECAP measures.

CHAPTER 4. QUALITATIVE STUDY OF THE CONTENT VALDIITY OF ICECAP-A: METHODS

4.1. Chapter introduction

This chapter reports the methods used in the assessment of the content validity of the ICECAP-A measure. The chapter begins with a description of the key characteristics of qualitative research and the methodological challenges associated with a qualitative assessment of the validity of a capability measure. This thesis used one-to-one interviews of informants who were purposively sampled in their professional role as researchers working in a health research setting. The selection of informants, the conduct of the interviews and the analysis of the data are described. The development of the *comparative direct approach,* a useful structure within which a thematic analysis of content validity can be completed, is described in detail.

4.2. Defining and explaining qualitative research

Counter to the mainly numerical evaluations that currently dominate validity assessment an emphasis is placed on qualitative methods as a way of assessing content and face validity [232] [265,266]. Rigorous qualitative methods are essential in the accurate assessment of content validity [233].

Qualitative research can be broadly defined as research that does not arrive at conclusions 'by statistical procedures or other means of quantification'$_{(p.17)}$[267]. Rather it attempts to 'grasp phenomena in some holistic way or to understand a phenomenon within its own context'$_{(p.171)}$[268]. To do this qualitative research often relies on language data, such as transcripts or reports, or observations to answer questions about "what?", "how?" or "why?"; rather than numerical data which quantifies "how many?" and "how much?" [269].

Qualitative research can be inductive, where patterns and associations are identified within the data and theories developed [269,270], or deductive, where theories or hypotheses are tested against the data [270]. In practice most qualitative research includes both inductive and deductive aspects [269]. It is hard to analyse data without existing theories or opinions influencing the results in some way while, equally, existing expectations or hypotheses will often not be fully formed and will need to be developed form the data [269].

The findings of qualitative research will 'inevitably be influenced by the interaction between the researcher and the researched'$_{(p.173)}$[268]. The researcher is normally placed within the research, allowing this interaction [238]. Unlike quantitative research which seeks objectivity, qualitative research has largely accepted that the values and opinions of the researcher play a role in the research. Rather than seeking to eliminate this, the role that researcher values play is examined and understood through the process of reflection [269], whereby the researcher self-consciously monitors their impact. Through this monitoring the researcher is able to assess the impact which their preconceptions, or theoretical assumptions, had upon the findings of the research [268].

Two broad categories of qualitative data can be defined: naturally occurring data, collected through participant observation and documentary analysis, and generated data [270]. Focus groups and one-to-one interviews are the most frequently used methods through which data are generated. One-to-one interviews can be used to form an in-depth understanding of a person's opinions, or ensure that a subject is covered in great detail [244]. Focus groups are normally used to form a broader understanding through examining the interactions between participants in a more natural setting than a one-to-one interview [244].

Samples for qualitative research should be selected in order to include the relevant constituents 'that can illuminate and inform'$_{(p.82)}$ understanding of a phenomena [271]. Sample recruitment, for both one-to-one interviews and focus groups can be purposive, where informants are recruited based on selected features or characteristics [271]. Purposive samples are often based on maximum variation sampling, whereby a sample is recruited based on widely varying characteristics, or theoretical sampling [239], where informants are recruited based on their ability to contribute to the goal of the research.

Unlike quantitative analysis of data, which has well-accepted procedures, qualitative data analysis has fewer agreed rules [241]. The approach taken to qualitative analysis will often vary with the characteristics of the data and the objectives of the research. While there are differences in the type of analysis used most qualitative analyses attempt to reduce voluminous and often unwieldy data into a more manageable form [272]. Types of analysis include: ethnographic accounts [273], conversation analysis [274], discourse analysis [275], analytic induction [276], framework analysis [269], thematic content analysis and grounded theory. It is the last two approaches which hold the most relevance to this thesis.

Thematic content analysis seeks to categorise recurring themes [269]. To achieve this transcripts from interviews or focus groups are read and re-read, and codes are used to classify segments of the transcript based, normally, on what the segment is referring to. The codes which are used can be developed through examination of early data to identify the key themes [269]. Once codes are assigned to the transcript the researcher can move beyond simple categorisation and begin to ask questions about how the themes relate to each other and why they may differ.

Glaser and Strauss [239,277] recommend that the analysis of qualitative data should use a constant comparative approach, whereby findings that emerge from the data are compared with previously collected data. Through a cyclical process of collecting and analysing data, and then allowing this analysis to inform future collection and analysis, a fuller account can be formed [269]. In grounded theory, like in thematic content analysis, transcribed data is analysed through the use of a coding structure. Glaser and Straus recommend that these codes are generated from themes identified in data and where possible take the form of words used by the informants themselves [239,240]. In grounded theory the application of the coding structure should be more intense than in a thematic analysis, with the goal of opening up or pulling apart the transcript for analysis [269]. Data collection should continue until saturation, where no new themes are being identified [269]. These key principles allow a thorough, inductive analysis of the data.

4.2.1. Methodological challenge

When measuring concepts such as quality of life, where a tangible concept is not being assessed, the questions of what is relevant and what should be measured are important [191,203]. This is a notable challenge for researchers attempting to validate the content of any measure. As described in Chapter 3, this methodological challenge can be addressed through the use of a nomological network: a scientific theory that gives a construct meaning through its links and associations with other constructs. However, when conducting qualitative research this network may have more limited use than in quantitative research. Qualitative research takes greater direction from the informant. Informants will have their own conceptualisations of quality of life (their own nomological networks), their own view of what content is relevant and what influences it. This will likely vary from informant to

informant. In a qualitative analysis that seeks to answer the question "does the measure sample the important and relevant dimensions of a construct?" variability, and a lack of clarity amongst informants about the important and relevant dimensions, are challenges, particularly in relation to concepts such as quality of life, where no universally accepted definition exists.

This thesis used one-to-one interviews of informants who were purposively sampled in their professional role as researchers working in a health research setting. A thematic analysis of the data was completed, which drew upon some aspects of grounded theory, particularly the analytic notion of constant comparison. In response to the methodological challenge identified, both interviews, and analysis, were structured in such a way as to allow a full understanding of the participant opinions of the ICECAP-A measure.

4.3. Informant selection and recruitment

4.3.1. Selection of informants

Informants were selected using purposive sampling, which allows informants to be selected based on certain characteristics [239,271]. Maximum variation sampling, a form of purposive sampling, allows a sample, which varies on selected characteristics, to be recruited [242], was used. This form of sampling can be useful in identifying the central themes across a number of different situations. Four groups were selected in order to provide a broad and representative range of opinions on quality of life and quality of life measurement from research professionals involved in controlled trials.

Informants were selected from four professional groups: 1) medical doctors involved in clinical trials and research; 2) clinical, primary care and public health trial experts, including

principal investigators and trial managers (this group will be referred to as "trialists"); 3) clinical research nurses or researchers with regular participant or patient contact (this group will be referred to as "frontline researchers"); and 4) health economists working within trials. Medical doctors and frontline researchers, who have regular patient and participant contact, have an outsiders – etic – view of how quality of life is affected by different diseases and conditions. In addition, frontline researchers have a unique perspective from regular participant contact while the participant is completing quality of life measures. It was hoped that they would be able to provide insight into how patients receive quality of life questionnaires. Trialists and medical doctors are involved in the planning and oversight of a trial. As part of the steering committee or management team of a trial they will select what measures are included within baseline and follow-up questionnaire packs. Health economists working with controlled trials advise on which quality of life measures should be used in the economic analysis. Between them these three groups determine which quality of life measures are used in trials.

In addition to these professional characteristics the participant selection sought to identify informants with and without prior experience or knowledge of the ICECAP measures. Furthermore, of those with prior experience of using, or knowledge of, the ICECAP measures, an effort was made to sample those with both a positive and a negative prior opinion. The recruitment of trials to the quantitative work of this thesis (described in Chapter 6) occurred before the qualitative research and identified a small number of researchers with negative opinions of the ICECAP measures and a number of researchers with positive opinions of the measures. These people were asked to participate in the qualitative work as informants and have been identified in Table 5 as having positive or negative opinions.

Informants working in both the UK and Australia were selected for interview. The ICECAP-A is being used in studies and trials in a number of countries outside the UK, Australia being one of these. The selection of Australian informants was designed to increase the generalisability of the research findings and identify any divergent views across the two countries.

Informant identification and selection occurred through a pragmatic process. Potential informants were identified through university and trial centre staff lists, as well as through existing researcher knowledge. The staff lists of the Medical Research Council (MRC) Hub Network for Trials Methodology Research, the Experimental Cancer Medicine Centre, Birmingham Clinical Trials Unit, and the Primary Care Trials Unit, at the University of Birmingham, were accessed online in order to identify suitable informants. Additionally, Prof. Joanna Coast and Dr. Hareth Al-Janabi were able to recommend suitable potential informants in the UK. Associate Prof. Paula Lorgelly and Prof. Stephen Jan recommended suitable potential informants in Melbourne and Sydney, Australia, respectively. The selection of Australian informants relied heavily on researcher knowledge, with all informants being identified through existing researcher knowledge.

Snowball sampling was used to identify additional informants. Snowball sampling involves asking informants, once they have been interviewed, to identify additional informants that fit the selection criteria [271]. Snowball sampling can be used to identify hard to reach groups. It was used here to identify additional frontline researchers, who often do not appear on trial organisation staff lists and were not known researchers on this project.

Recruitment of informants was stopped after data saturation, the point at which no new themes were being identified in the data [269], was reached on the primary objective of the research (the validation of the ICECAP-A measure). Data saturation on emerging themes, not central to this objective, was not targeted.

4.3.2. Recruitment of informants

Potential informants were approached through a standardised e-mail (Appendix 5) describing the purpose of the research and asking for the cooperation of the potential informant. An information sheet was attached to the email (Appendix 6) and the ethical approval gained for the study (Appendix 7) was explained. A follow-up phone call was scheduled for five to eight days after the email was sent, in the event of no reply. On receipt of a positive reply a date, time and place was agreed for the interview. If the date set was more than two weeks in advance then a reminder email was sent 2-5 days before the interview. All but one of the interviews were conducted at the informants place of work (one interview was completed at a university café).

4.4. Interview conduct

At the start of the interview informants were asked to sign a consent form. The confidential nature of the interview and the fact that the interview was being recorded using a digital voice recorder was emphasised to the informant verbally at the start of the interview.

The interviews were semi-structured and designed to allow breadth and depth of discussion, while investigating some areas in greater detail. A topic guide (Appendix 8) was used as an *aide memoire* to encourage a consistent structure to the interviews and adequate coverage of

key topics. This topic guide was flexible and updated throughout the interview process to allow development of emerging themes. Interviews were broadly partitioned into two areas.

4.4.1. Interview part one

The first part of the interview assessed the informant's understanding of quality of life as a concept and their own conceptualisation of quality of life, with particular focus on their research areas. This was done in order to add context and increase understanding of data gained in the second part of the interviews. The focus of this initial part of the interview was to encourage breadth and depth of discussion and allow informant to define, as fully as they were able, quality of life and its influences. Informants were also encouraged to give their opinions on quality of life measures they had previously used in their research. This was designed to enable some understanding about the informants' general opinion of quality of life measures.

Content mapping was used. Content mapping includes the use of ground mapping, questions to open up the subject area; dimension mapping, questions to focus the informant on a key topic; and perspective widening, questions to widen the informants perspective [278]. Ground mapping questions were used to encourage the informant to describe their conceptualisation of quality of life.

"If I said Quality of Life to you, what would you take that term to mean?"

In order to ensure that informants provided more than their first thoughts or immediate reaction the informant was encouraged to take time to think about their answer. To facilitate this, perspective-widening questions were used.

> *"Now you have described health as a major influence on quality of life. Are there other factors that you think may influence someone's quality of life? [pause] Take your time."*

Where participants struggled to provide a full description in the level of depth required, content mining questions were used to allow the informant to explore further and explain more clearly the concepts they defined.

> *"You said contact with family is important. How would this influence someone's quality of life?"*

Conversations in qualitative interviews are rarely sequential and numerous points of interest are often raised in quick succession or out of context, or when they are raised are often poorly defined. Dimension mapping questions and clarifying probes were used to ensure that the meaning of the informant had been understood and was not assumed.

> *"So can I check I understand, you feel that quality of life is influenced to a large degree by health, but other factors such as friendship and independence are also major influences. Is that correct?"*

4.4.2. Interview part two

In the second part of the interview measures were introduced. Informants were encouraged to discuss at length the ICECAP-A and EQ-5D-5L measures and provide their opinions on the measure's content and their view of its usefulness to their research area. Interview informants were presented with copies of the ICECAP-A and EQ-5D-5L one after the other. The order in which the measures were shown to the informants was random except in the situation when it was beneficial to the flow of the interview to consider one first. For example: if towards the end of the first part of the interview the informant was discussing concerns with the sensitivity of the EQ-5D-3L, the EQ-5D-5L would be presented to the informant first. Informants were encouraged to discuss the measures at length.

The appropriateness and usefulness of the measures for use in the informant's research area were discussed. Content mining questions were used to understand decisions about use of measures.

> *"So thinking about the research area you work in and the patients you work with, do you think that this measure is suitable for use?"*

When an informant found making a comparison between the two measures useful in describing their opinions of each measure, this comparison was encouraged. However, such comparisons were not prompted by the facilitator.

The content validity of the EQ-5D-5L was examined in this work. The results of this analysis are not presented in the main body of this thesis, but can be seen in Appendix 10

4.5. The comparative direct approach

The *comparative direct approach* was developed based on researcher experience from early interviews and through using data and experiences from the first batch of data analysis (the first 4 interviews). This approach attempts to address the methodological challenge of content validation identified in section 4.2.1. Methodological challenge, by providing a useful structure with in which a thematic analysis, grounded in the data, could be completed. The process of development and the final approach used is described below.

4.5.1. The development of a method

A two-part interview was used from the outset. This two-part partitioning of the interview provided an opportunity for the development of an innovative two stage approach to assessing content validity. The analysis of the first batch of interviews highlighted the possibility of

identifying similarities between the informants' description of quality of life and its determinants, and the descriptive system of the measures under consideration. The researcher was able to identify when the informant was discussing the importance of a dimension which was assessed by the measures. As an example take the three brief passages from the same informant below. In these examples the informant describes issues relating to three of the dimensions of the EQ-5D-5L descriptive system as influences on quality of life: A) Pain, B) Mobility, C) Usual Activities.

> *A) "…well obviously there is a pain thing, so the higher the pain the lower your quality of life"*
>
> *B) "…I mean you could have someone with a bilateral amputation below the knee who's in a wheel chair and their quality of life is zero"*
>
> *C) "Work...that's a BIG thing really"*

Early analysis also showed potential to identify dimensions that informants felt were important, but were not included in the descriptive systems of the measures under consideration. In the example below the same informant, talks about the important of communication and sensory ability, which is not directly assessed in the EQ-5D-5L.

> *Being able to see and hear. Also, speak to people. Interaction with the world is all about this.*

The discussion in the first part of the interview normally had a broad scope, which was notably reduced in the second part of the interviews, when the measure was considered directly. During the first batch of interviews it quickly became apparent that this initial discussion was not only useful as a reference point to understanding the later opinions of the informant, but also in assessing the content of the measure in itself.

In the second part of the interviews informants were encouraged to directly give their opinions of the measures under consideration. Content validity was examined through discussion of whether informants felt the content of the measure covered the important and relevant dimensions of quality of life. Face validity was assessed through discussion points such as whether the length was appropriate, whether it was understandable and if it would be of use in the research area in which they worked.

In the early interviews, when discussing the content of the measure, informants frequently referred back to *their* earlier description of quality of life, and discussed whether the measure did or did not offer coverage of *their* definition. It was noted early in the analysis that informants gave full and expansive discussion of the measures when making this comparison. This was identified as a rich data source. The topic guide was altered for later interviews to include questions and prompts that directly encouraged the informant to refer back to the earlier discussion. The informant was encouraged to think about their previous description of quality of life and its determinants. They were asked to discuss whether the measure under consideration covered what they felt quality of life to be. The informant was also encouraged to consider each items each measure individually and comment on relevance and completeness.

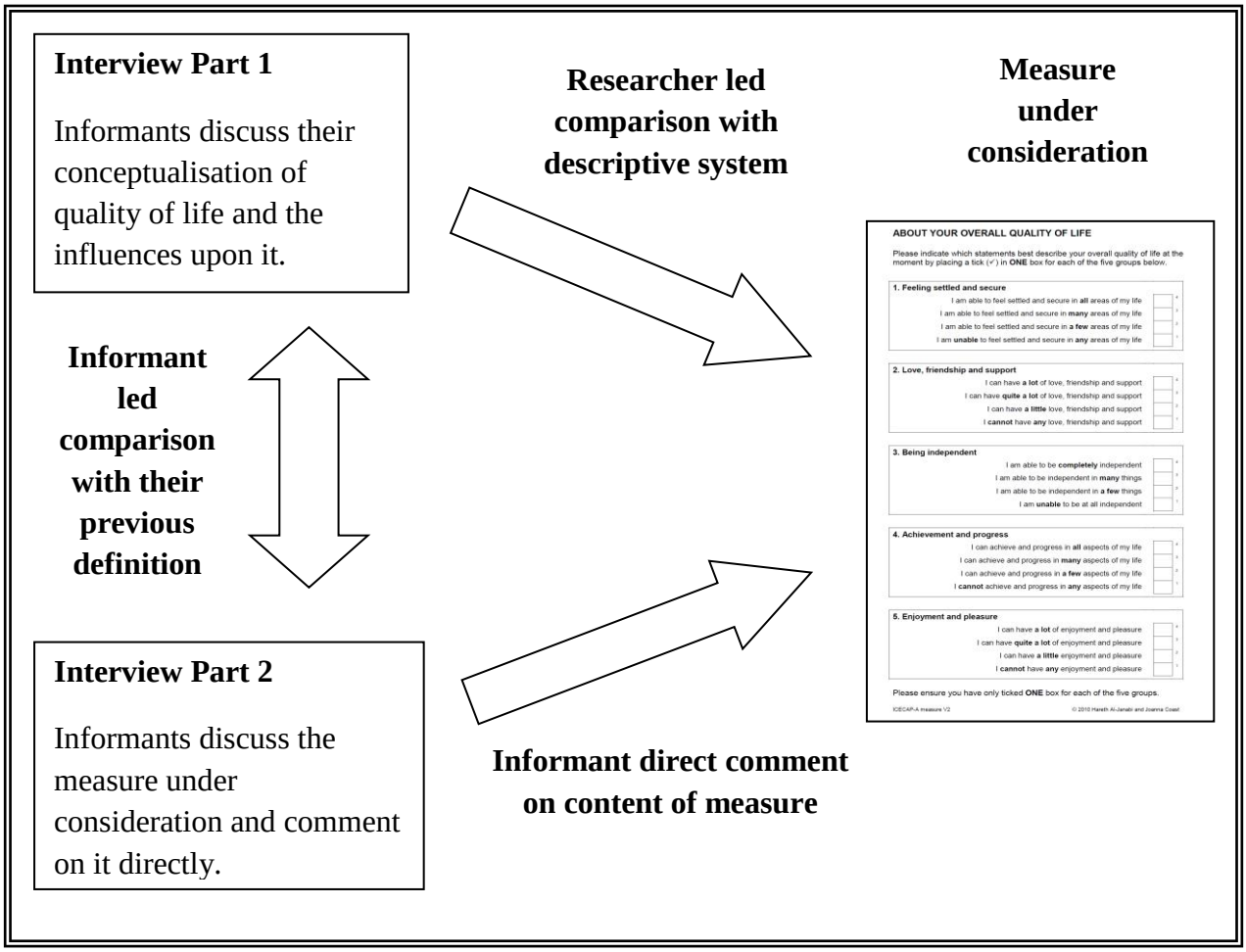
> *"So thinking back to how you described quality of life, and what influenced it, do you think this measure covers that?"*
>
> *"Now if we could go through the measure and consider the content of each item individually"*

4.5.2. The comparative direct approach

These experiences from the early interviews led to the development of an approach that allowed a fuller assessment of the content validity of the EQ-5D-5L and ICECAP A measures

than would have been possible at the start of the interview process: the comparative direct approach. This is an effort to respond to calls by Magasi and colleagues [233], Brod and colleagues [232] and others for rigorous and transparent qualitative methodology in the assessment of content validity or patient reported outcome measures.

Figure 9: Graphical presentation of the Comparative Direct approach.



The first stage of the analysis, the *comparative* stage, used data from the initial part of the interview, where the informant defined their understanding of quality of life as a concept. This description of quality of life is *compared* by the researcher with the content of the measure under consideration. This part of the interview has the advantage that the informant has not been influenced by seeing the measure. This provides a reference point for the

analysis for the informant; they have described what they define quality of life to be and what influences it.

The second stage, the *direct* stage, uses data from the second part of the interview where the informant has the measures in front of them. The informant discussed the relevance of the measure and the overall content coverage of the measure as a whole and each item in turn. In this part of the interview the informant was asked to think back to what they defined quality of life to be in the initial part of the interview, and assess whether they felt the measure covers their conceptualisation of quality of life. Using the reference point established earlier in the interview is important. It addresses the methodological issue of there being no widely accepted definition of quality of life, by allowing understanding of the informants conceptualisation, and allows a better understanding of the informants' opinions on the coverage of the measure.

4.6. Data handling

Interviews were transcribed verbatim using Olympus transcription software. All names and references to locations that would have indicated the informant's identity were removed. Each informant was assigned a code. An electronic index of the codes and the corresponding informant was kept in a password protected document. All electronic transcripts were stored on a password protected computer in a locked room. Hard copies of transcripts were kept in a locked cabinet in a locked room.

4.7. Data analysis

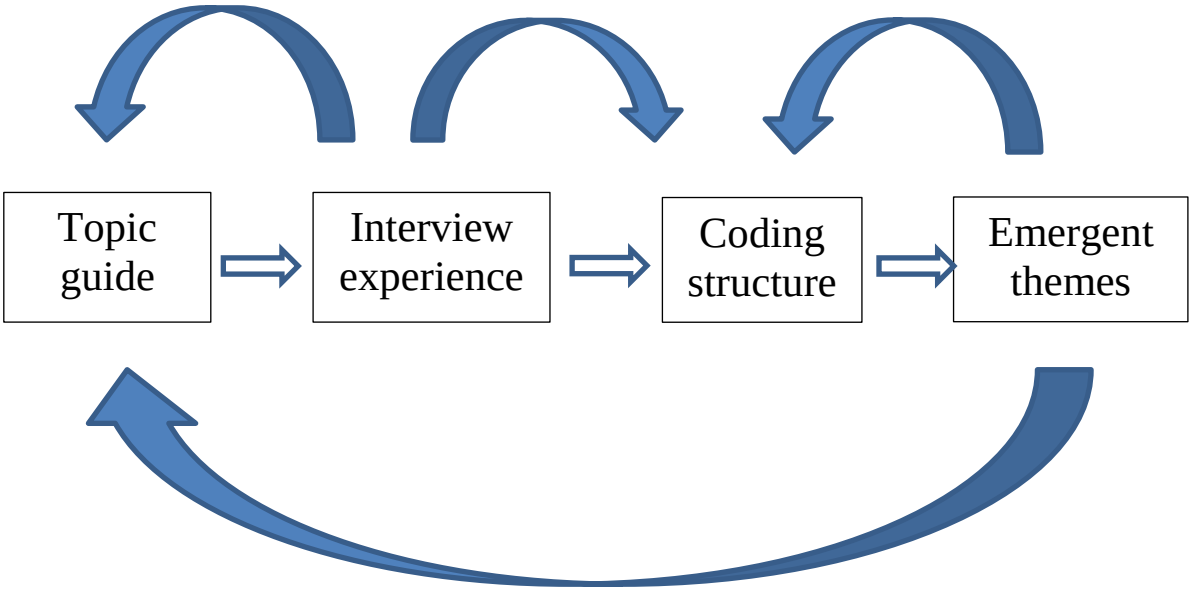
Interview transcripts were coded using the Atlas-Ti computer-assisted qualitative data analysis software. A hierarchical coding structure was formed during the analysis of the first batch of interviews and was grounded in the data and researcher experience of the early interviews. Where appropriate the participants' phrases and language was used to form codes. The coding structure was updated and refined throughout the subsequent batches of analysis in an iterative manner, to allow the identification and development of themes. The final version of this is placed in Appendix 9. The process of updating and refining the coding structure is presented in Figure 10.

An iterative, constant comparative, thematic analysis of the transcripts was completed. Transcripts were analysed in four successive batches. This analysis allowed descriptive accounts to be formed. These accounts were formed for each batch of the analysis and sought to synthesise the data, highlight key trends and map the diversity of opinion amongst the informants [241]. The iterative nature of the analysis allowed themes which were identified in earlier batches to be analysed and developed in later batches of the analysis. Emerging themes were initially identified and then discussed with doctoral supervisors, before being assessed in greater detail both through refinement of the interview topic guide and greater focus during analysis through the use of the flexible coding structure (above).

At the conclusion of the work an explanatory account was formed, which through assessing patterns within the data and drawing comparisons between informants, sought to move beyond descriptive analysis to explanatory analysis [241]. A focus was maintained both on participant opinions of the ICECAP-A measure, but also on understanding the reasons for the

opinions held. This explanatory account formed the basis from which the following chapter was written.

Figure 10: A representation of the iterative nature of the thematic qualitative analysis



Solid fill lines indicate the refinement of the topic guide and coding structure based on interview experience and emergent themes.

CHAPTER 5. QUALITATIVE STUDY OF THE CONTENT VALIDITY OF ICECAP-A: RESULTS

Note on chapter: The qualitative results presented in this chapter have been published. A copy of this publication is contained in Appendix 10. The reference for this publication is:

Keeley T, Al-Janabi H, Lorgelly P, Coast J (2013) A qualitative assessment of the content validity of the ICECAP-A and EQ-5D-5L and their appropriateness for use in health research. *PlosOne.* 8;12.

5.1. Introduction

This chapter reports an assessment of the content and face validity of the ICECAP-A measure using rigorous qualitative methodology described in Chapters 3 and the methods described in Chapter 4. The development of the two stage comparative direct approach provided a useful structure in which this analysis could be completed. The reporting of results in this chapter reflects this structure. First, the informants' conceptualisation of quality of life, which was discussed in the first part of the interview, is reported. The chapter then continues to discuss the informants' perception of the ICECAP-A measure. Within this section data referring to informants' overall perceptions of the measure is presented first before an item by item breakdown of the data is completed. In the item by item analysis comments before and after viewing the measure are presented and compared.

An emergent theme of how quality of life measures are chosen for use in randomised controlled trials is reported. A conceptual model is offered based on this data. Discussion of these findings in the context of the ICECAP measures is provided.

5.2. Interview recruitment and informant characteristics

Seven male (IS03, IS10, IS11, IS12, IS13, IS16, IS19) and 10 female research professionals were interviewed between January and September 2012. Eight interviews were conducted in various locations in Australia; nine interviews were conducted in various locations in the UK.

Informants were selected from four broad professional roles and different clinical areas, shown in Table 5. Four informants were classified as "frontline" researchers. Two frontline researchers were trained as nurses (IS01, IS05), one as a physiotherapist (IS09) and IS17 had no clinical training. Seven informants were classified as trialists. The group included principal investigators (IS16, IS06) and senior research fellows (IS04, IS07); all those in the "trialist" category were experienced investigators in terms of length of time in research and number of projects worked upon. Two "trialists" were clinically trained (IS04, IS16), but were not working within a trial as doctors; two had some knowledge of health economics in clinical trials (IS16, IS10). Three informants were classified as research doctors, meaning their primary role in a trial was a clinical one. They have therefore been classified under the "trialist" category. Three health economists were interviewed.

All but two (IS04, IS11) informants had experience of using the EQ-5D-3L (see Table 5). Five informants had experience of using or had previously seen one or both of the ICECAP measures, with three (IS01, IS16, IS17) informants holding an *a priori* positive view and two (IS02, IS12) informants holding a negative view of the measure. None of these changed their views after considering the ICECAP-A measure in the interview suggesting the views they held on the ICECAP-A measure were already well established. A number of other measures

had been used by the informants, the most common being SF-36 and the EORTC (see Table 5).

Informants worked in a number of clinical areas (see Table 5) across primary (IS06, IS07, IS09, IS10, IS12, IS16, IS17, IS18, IS19) and secondary (IS01, IS02, IS03, IS04, IS05, IS08, IS11, IS13) care settings.

Data saturation identified as having occurred by interview 14. This was judged as the point at which no new themes were being identified in the data on the primary objective of the research (the content and face validation of the ICECAP-A measure). Data saturation was not reached on the emerging theme of how measures are selected for use. Three additional interviews were conducted to check saturation had occurred and to ensure adequate numbers were sampled from each professional role.

Table 5: Professional characteristics of informants included in qualitative research

Informant	Location	Primary clinical area	Role	ICECAP-A prior opinion	QoL measures used
IS01	UK	Cancer	Frontline	Positive	EQ-5D EORTC ICECAP-O
IS02	UK	Cancer	Trialist	Negative	EQ-5D EORTC
IS03	UK	Cancer	Trialist	None	EQ-5D QLQ-30 PEDS-QOL
IS04	AUS	Arthritis	Doctor	None	SF-36 SF-12
IS05*	UK	Cancer	Frontline	None	EQ-5D SF-36
IS06	AUS	Physiotherapy	Trialist	None	EQ-5D SF-12 SF-6D
IS07	AUS	Injury Prevention	Trialist	None	EQ-5D
IS08	AUS	Cancer	Health Economist	None	EQ-5D EORTC
IS09*	AUS	Physiotherapy	Frontline	None	EQ-5D SF-12
IS10	AUS	Multiple areas	Trialist	None	EQ-5D SF-36 AQoL
IS11	AUS	HIV	Doctor	None	AQoL HIV spec
IS12	UK	Primary Care	Trialist	Negative	EQ-5D SF-36
IS13	AUS	Blood pressure	Doctor	None	EQ-5D SF-36
IS16	UK	Public Health	Trialist	Positive	EQ-5D SF-12 SF-36 AQoL ICECAP
IS17*	UK	Stroke	Frontline	Positive	EQ-5D ICECAP SF-36
IS18	UK	Primary Care	Health Economist	None	EQ-5D SF-36 SF-12
IS19	UK	Pharmacology	Health Economist	None	EQ-5D

* Informants who were recruited through snowball sampling. Please note: due to a coding error the codes IS14 and IS15 were not assigned to informants.

5.3. Content and face validation

This section presents evidence of a thematic analysis of the content and face validity of the ICECAP-A measure. The informants' conceptualisation of quality of life and its determinants are described. The direct comments made by informants upon the content of the measure and an item-by-item content analysis of the ICECAP-A using the comparative direct approach follows. The relevance and appropriateness for use of the measure in the informants' research area are detailed.

5.3.1. Quality of life beliefs

5.3.1.1. Quality of life as a broad concept

A strong theme running through the interviews was that informants perceived quality of life to be a broad concept. A common initial response to some variation of the probe "what do you hold quality of life to be?", was for informants to state that they thought it was a broad construct. Quality of life was frequently referred to as "big picture", "multi-dimensional" or "broad".

> *IS16 So it has got to be all things to all people.*
>
> *IS06 So I guess I view it as a broad construct which is influenced by all of the other aspects.*
>
> *IS12 it is obviously multi-dimensional*
>
> *IS16 I think it is a sort of "how are you" question.*
>
> *IS09 and I think...it has to be sort of big picture*

Informants discussion of what determined and influenced this broad construct of quality of life could be categorised under three main themes, with each theme having a number of sub-

themes: physical health and impairment, which included ill health, pain, functioning, mobility and treatment side effects; psychological state, which included psychological health and outlook on life; and social, life and living, which included family and friends, work and ability to lead a normal life.

5.3.1.2. Physical health and quality of life

Physical health was recognised as being an important determinant of an individual's quality of life. A high level of agreement existed amongst informants that poor physical health or physical disability reduced quality of life.

> *IS09 I still see health as important in that [quality of life]. I think when someone has got ill health...it is quite a big determinant.*

Informants were able to give a number of examples of how specific illnesses or conditions can reduce quality of life. Informants used examples from their area of research as well as personal example or examples of extreme health states, which emphasised their point.

> *IS05 ...you could have somebody, bilateral amputation below, knee who's in a wheel chair and their quality of life is zero.*

> *IS13 To be incontinent, of either urine or faeces, you know, by God your quality of life goes down. So there are all sorts of terrifying things which can happen.*

Pain was a category of health that informants frequently identified and placed particular emphasis upon. Pain was discussed at greater length by informants who had regular contact with patients or trial participants (frontline researchers and research doctors), who noted the pervasive influence it can have on quality of life.

> *IS18 So you could have chronic back pain and…that would certainly, I mean yeah poor health impact on quality of life, there is no doubt about it.*

> *IS13 Nothing worse for quality of life in many ways than chronic discomfort and pain.*

For informants who worked in cancer research, the physical side-effects of treatment were identified as having a notable influence on quality of life. There was a consistency amongst these informants in the way side-effects were described as affecting patients, mainly through fatigue. No informants from any other clinical area discussed side-effects.

> *IS08 Treatment is probably a predictor of the quality of life. So the chemotherapy is of course and radiotherapy both have a high impact on fatigue.*
>
> *IS05 They could be perfectly well and asymptomatic of their disease and then us give them lovely cytotoxic chemotherapy and a bit of radioactive isotopic one off injection which reduces their quality of life completely.*
>
> *IS01 Some of the new wonder drugs that have come through, [we] now realise have huge cardio-toxicities attached to them…which is very life limiting and well-being limiting.*

Mobility and physical functioning was viewed as allowing a person to complete their daily routine and maintain independence. Mobility and the ability to physically function was valued not just for itself, but for what it allowed a person to do. This is closely linked to the normal activities section below. Informants who worked in physical therapy and falls prevention frequently discussed this point.

> *IS02 ...well anything that would impair your ability to live a normal life. So it could be physical in that you can't walk or paralysis or something extreme like that. Or it could be pain that could make it difficult to function.*
>
> *IS03 Yeah, well there is physical functioning...can you live a normal as it were life? Can you deal with day-to-day things?*
>
> *IS04 …they can't get down to the shops to do their shopping. Or, they want to do something and they can't do it. It is too hard, they have got to think about is it feasible to do something that they want to do based on how mobile they are. It limits your options.*
>
> *IS06 …physically it might be catching a bus to see their daughter or going down to the shops or gardening or something like that.*

Health was considered a major influence on quality of life. However, informants indicated that they did not feel that it was an overriding influence and were of the opinion that those in poor health states were still able to have a good quality of life. When discussing this point a number of informants used personal examples of friends and relatives who were living with, or who had lived with, illness or disability.

> *IS16 All you see is ill health and states that you don't want to get into, but there are people that get into those states and have a fantastic time.*
>
> *IS01 But no he certainly has good quality of life, but he's got chronic pain and breathlessness and fatigue; from the outside point of view it's harsh, but by no means depressed or fed up. Still the life of the party. Just not for as long and not with as much red wine. He has shortened some of the things but there is still quality there.*
>
> *IS09 ...some people seem to be able to cope with a remarkable amount and yet still be able to...access the social contact or be involved in the things they want to do.*

Much of the discussion about people in poor health being able to maintain their quality of life, was focused around the ability to cope with and adjust to their health state. This process of adjusting or adapting was discussed from both a shift in perspective of the individual and practical adaptation of learning to live with the condition.

> *IS16 So then they are not limited and they see it as their lot [in life], and they get on with life.*
>
> *IS06 So basically they will learn to walk with a prosthesis or learn to be independent using a wheel chair. ...so then their mobility will obviously be affected to some extent, but they can still still you know get back to doing things that they enjoy doing.*

5.3.1.3. Psychological health and quality of life

Informants thought psychological health was an important determinant of quality of life. The majority of informants interviewed discussed the effect of depression or emotional problems, and all those who discussed if felt that it was an important influence on someone's quality of life.

IS18 it is so pervasive into everyday life, probably almost more than anything else.

IS18 yeah, well I think mental health would impact very strongly.

Informants discussed mental health as a distinct, separate category from physical health.

IS12 I can I am sure there are people that are fantastically physically fit but have a terrible quality of life because they just are at odds with the world.

While this distinction was made between physical and psychological health, a link between the two was evident. Informants were of the opinion that psychological ill health, such as depression, could be a result of physical ill health.

IS13 if your illness is causing you more and more trouble you could end up getting perfectly understandable amount of depression. Depression is a frequent co-morbidity of severe physical illness.

An informant discussed how the internalised perception of being a "patient", even in the absence of physical symptoms, may affect a person's quality of life.

IS16 They aren't feeling as well as they could be because they recently diagnosed as hypertensive and they are not unwell with that, but they are not they no longer consider themselves as well, because they are now having to take a drug every morning. And remembering that. And that is the difficulty that quality of life faces as a term.

A second theme under the overarching psychological topic was psychological outlook or outlook on life. This was distinct from psychological health discussed above, as it was not viewed as an illness or condition. Informants discussed psychological outlook as a mental attitude to life or a psychological frame, which might determine an individual's response to situations.

IS09 I think that some people seem to have been very resilient and get through things and other people get bogged down.

IS13 It depends a lot on, it depends a lot on mental attitude. So you can have people that are seeming to have all the rotten luck in the world and think their lot in life isn't bad. While you have other bastards that have everything on a silver spoon and think

their lot in life is shit.

Psychological outlook was perceived as having an interaction with physical health, which resulted in different people reacting differently to illness and disability. Those with a positive outlook on life were thought to be able to deal with personal ill health better than others.

> *IS16 I think health affects people in very different ways depending on what the condition is, but also their own psychological frame and their own ability and the resources around them to support them through that so, plenty of people in the world, who inevitably think the worst of anything that is diagnosed, and that inevitably leads to a certain sort of mindset and action.*

5.3.1.4. Social, life and living

When discussing the topic of social, life and living a greater use of scoping and content mining probes were required. A number of informants raised this theme, but in comparison to the ease which all informants discussed physical and psychological health (above), informants showed a greater level of struggle when discussing this topic. While informants struggled to discuss this theme, it did not appear that this struggle was linked to the importance they attached to it. Rather it appears that informants had greater trouble verbalising their thoughts. The use of appropriate probes, which brought greater structure to this section of the interview, allowed informants to discuss social, life and living in depth. Taken as a whole the data suggest these informants felt there is a broad spectrum of social influences on a person's quality of life.

> *IS13 things that can impact on a person's life other than being in your trial. Life is going on, having children, they are getting married, they are dying, they are living, having operations, they're going broke, they are making a fortune, and whatever else is happening.*

Family and friends were identified as an important determinant of quality of life. Informants felt that as well as the enjoyment that people get from spending time with them, family and

friends provide support and assistance which can increase the quality of someone's life. Examples were given of people who were in poor health, finding enjoyment from family and friends.

> *IS04 the computer keeps his world open enough that people five years later still come to visit every day. People who are not necessarily family, people who were friends, people he didn't really know before...who started coming and then keep coming. It is pretty impressive.*
>
> *IS01 Thinking of a family member who has considerable impact on his health from lung cancer, but when he is around his family, doing the things he enjoys.*

The situation where someone loses a loved one was frequently used to highlight the importance of friends and family.

> *IS06 … any major life events, particularly with older people, if their spouse has died, or their daughter's not helping them, or something like that, probably has an influence*

A family support structure was noted as important for those in ill health. The majority of discussion on this topic was focused on the situation where a support structure is absent and the negative impact this can have on an individual.

> IS05 *"gosh this poor family"…his wife couldn't deal with it, he couldn't deal with it, they weren't working together, they weren't thinking about the other person. They had the daughter there who was just a mess and his wife actually turned around to me and said "how to I support him"…. And, his quality of life it wasn't just wasn't the cancer, it was the fact his family support structure, because it wasn't there.*

Work and financial security was mentioned by a number of different informants. Work was viewed as positive and a negative perspective influence on quality of life. Informants discussed how work can be an enhancing factor in someone's life, and the example of Stephen Hawking was used by several informants.

> *IS08 I think my quality of life in Australia for example, living here, and it includes the weather and it includes what type of job I have and how healthy I am at the same time.*

IS02 you could think about Stephen Hawking, mentioned a lot at the moment. But he is physically very very seriously impaired and yet is he 70 now? With a very, what must be a very, satisfying rewarding job you would think.

A small number of informants noted that the stress and pressure associated with work can limit life; however this was in contrast to the predominant view of work and ability to work being a positive influence on quality of life.

IS01 work as well, being able to get the time away from work without feeling the pressure of the place.

The ability of people to lead their normal life was a strong theme. This topic was often discussed with reference to how loss or reduction of health affected quality of life. This section was closely linked to health and mobility (above), and indicated that these things were not just valued for themselves, but rather what they allowed.

This discussion focused on the ability to do things that they normally do, or lead a normal life. When pressed on the meaning of a normal life, informants were quick to explain that they didn't mean a generic or standardised life, rather the life that the person had or wanted to live.

IS06 So it could be in terms of their participation, that they are not able to do the things that they normally do. So whether it is looking after grandchildren or cooking meals or something like that. So that aspect of their life can be affected. That they perform social roles that they would normally perform I guess.

IS09 I suppose I would think [quality of life] that someone is doing what they want to be able to do.

Within this discussion was the need for individuals to achieve things of value. The examples given by informants were not of big, one off achievements (such as running marathons or climbing mountains), rather everyday day actions that defined the individuals or filled their day-to-day life.

IS12 …he was in and out of hospital, he was in a lot of pain, he couldn't do all of the things prior to that which had great meaning and value to him. He used to keep pigs

and chickens, you know small holding, a really huge part of his life he had to give it all up he couldn't do it. Too ill.

5.3.2. Informant opinions of the ICECAP-A measure

Informants viewed the ICECAP-A measure as a broad assessment of quality of life, generally appropriate for use in the research fields in which they worked. The measure was viewed as a short, uncomplicated measure, suitable for use with participants in a busy research environment:

IS01 It is a lovely length...because they haven't, you don't have the time to spend with a long questionnaire.

IS18 I think that is very clear and it is obviously quite straight forward in that sense to actually physically fill it in. Yeah.

Informants noted that it was different to existing health-related quality of life measures, focusing on the emotional determinants of quality of life.

IS06 *Yeah I guess this one is more general...and focuses mostly on the emotional.*

IS07 Its very much the emotional aspects of things.

When asked to think back to how they defined quality of life in the first part of the interview and comment on whether the ICECAP measure captured that definition, the majority of informants felt that it did.

IS06 And, to me this is more about psychological aspects. Yeh which I think probably is actually measuring overall quality of life.

IS18 So would that capture my quality of life [long pause]? Errrr [long pause] yyyeessss pretty much, pretty much.

The primary concern that existed was that the ICECAP-A did not directly assess health, which informants had identified as an important influence on quality of life in the early part of the interview. A number of informants noted that this. Some felt this was an important

dimension that the measure did not assess, which reduced the relevance of the measure to their clinical area.

> *IS19 It is just how far away from health it gets I suppose...I think it is just the distance from health [which is a concern].*
>
> *IS18 a lot of them are health-related quality of life…this clearly isn't.*
>
> *IS04 Yeah, no, that [ICECAP-A] captures Dad's quality of life and this is the one [EQ-5D-5L] we use in studies.*

Several informants made comparison between the EQ-5D-5L and the ICECAP-A measure. While different informants had preferences for different measures, these comparisons suggested that informants tended to view the measures as measuring two different things. The EQ-5D-5L was generally viewed as measuring health; whereas the ICECAP-A was viewed as capturing a more general definition of quality of life.

> *IS13 when you look at them...this one strikes me again as quality of life, making a judgement that emotional things are more important, because that is more emotional...wellbeing and this [EQ-5D-5L] is more total health.*
>
> *IS18 So this [EQ-5D-5L] is really about THE person and their little box situation, whereas this [ICECAP-A] is much broader about their life in general.*

A tension was present in the data between informants who viewed the measure as assessing dimensions that matter to patients and a small minority who felt the subject matter were too sensitive and not appropriate to ask patients. The perception of inappropriateness was motivated by a concern over the questions being upsetting for people who had low levels of the dimensions being assessed, rather than being inappropriate *per se*. Informants who felt the measure was patient-focused were more likely to be frontline researchers or research doctors, while those who held concerns over the measure were more likely to be trialists.

> *IS01 I think they would be [happy to answer], because it sums up the kind of conversations you have with patients and I think they would be quite comforted with it.*

IS18 I don't think they would object. I don't think anybody could object to filling that is.

IS02 I find them [the questions] less appropriate... If you were stuck in a dead end job might be cheesed off by asking if you can achieve and progress because you spend most of your working days doing something that is soul destroyingly dull.

The majority view of informants was that the measure was appropriate for use in the research area in which they worked, with only a small number providing a contrasting opinion.

IS11 I like it. I love it. That is my initial feeling. I love it.

IS09 No I quite like that. I think it covers a lot of issues very simply.
TK And do you think it would be of use in the research areas that you work in?
IS09 Yeah I do. I think this is a good basis for knowing where someone is at.

A small number of informants commented that it might be of greater use for research with the general population. These informants noted that for a population of very sick people it may be a bit too broad for use.

IS10 But it depends, if you are doing a broad research setting, not every research has to be done in people who are really sick. It could be a very broad intervention...that could be very applicable. But if it's you know if it's an oncology and people are really sick and people are going to die, that's probably a bit too broad.

There was a consensus that the measure would be best used in addition to, rather than as a replacement for, existing health-related quality of life measures. Informants felt that something that provided more information about the source of the problems and maintained a focus on health-related quality of life was also required. This was motivated by a perception that the ICECAP-A measure was not measuring health-related quality of life.

IS18 I think it would definitely [be of use]...I probably wouldn't replace the EQ-5D but I would think quite seriously about using it in addition.

IS19 So it would complement, I wouldn't see it as a replacement for an EQ-5D, but it would certainly complement an EQ-5D type instrument.

A number of informants discussed the capability wording of the ICECAP-A measure. Some informants showed a level of cognitive struggle in understanding the focus of the question and a level of concern existed about whether the wording would be understandable for participants in the studies.

> *IS06 Yeah I guess the "can" and….the difference between can and do maybe might be confusing to some people…it might be clearer to leave out the can.*

However, in the majority of the discussions about the capability wording informants reached an understanding that would be broadly in line with the capability theory (i.e. they understood the question correctly).

> *IS03 I don't like the "I am able" or "I can", you know I don't know, it feels as if in some way you are the person with the control , so I CAN have a lot if I want to I can have a lot of love and friendship.*

5.3.2.1. Item by item analysis

Below is an item by item analysis of the ICECAP-A measure whereby the comments before and after viewing the measure are presented and compared. A description of what the item is designed to measure is taken from Al-Janabi et al [115] and is provided at the start of each item. Many of the comments provided after viewing the measure were elicited through asking the informant to think back to how they had previously described quality of life.

5.3.2.1.1 Stability

The Stability item was designed to assess an individual's ability to live a life of continuity, without feeling concern or uncertainty [115]. Prior to viewing the measure, informants identified stability as an important dimension of quality of life. Living with fear and uncertainty due to a physical condition or illness and the concern that unemployment due to illness can cause, was identified by a number of informants as a pertinent issue.

IS13 You get frightened of taking your medicine. You get frightened of going to sleep, in case you don't wake up.

IS08 ...getting back to work. The worries of losing a job, not getting back, not being able to get jobs back.

Upon considering the measure, there was a broad acceptance that the Stability item was relevant to the assessment of quality of life. Informants indicated that the item would resonate with participants of different ages, and would be influenced by both health and non-health factors. A number of informants recognised that the item was assessing a construct that they had identified as important in the earlier part of the interview.

IS01 ...it makes sense because...the phase one patients I see are very palliative and they don't have a lot of time. But you can still be settled and secure with months to live.

5.3.2.1.2. Attachment

The Attachment item is designed to assess the importance of support, social contact and relationships [115]. Prior to considering the measure, informants identified the ability to function in a social context as an important consideration. Relationships were identified as important both for the enjoyment and support they provide. A loved one dying was often given as an example of the importance that relationships can have on quality of life. The importance of people suffering from illnesses to achieve social contact and the limiting effect that illnesses can have upon one's ability to achieve social contact was discussed at length.

IS12 And in the last year of his life, he died by the cancer, he said...this has been the best year of my life, because until this moment I never realised how loved I've been.

IS13 ...people who are incontinent tend to become completely reclusive, they don't want to mix, they are terrified. They don't want to go out there and pee in front of everyone.

Upon considering the measure informants recognised Attachment as a relevant item to ask. It was also noted that it is an area that is not often assessed.

> *IS11 Well things like love, friendship and support. It is all that thing around social connectiveness and support and intimacy. We as a research group are very interested in that in people with HIV.*

A level of concern was expressed by a number of informants, over two particular points. Firstly, there was a concern that the item was assessing a number of different concepts (love, friendship and support) within one question. The possibility that a person may have one, but not all of the concepts being assessed was a concern.

> *IS03 ...I could just imagine, particularly with love friendship and support, I could imagine someone saying they have got one but not the other things. Might be difficult.*

A second point of concern was the sense a few informants had, that this was a sensitive subject. These informants were trialists, without regular patient contact. This concern focused on the prospect of asking the question to an individual who had recently lost a loved one or in the process of a relationship breakup. Some informants who had identified Attachment as important before viewing the measure had this concern.

> *IS02 I don't think number two is very appropriate. There is a person here, they may, they haven't got any love or friendship. That is something that is completely outside their control.*

5.3.2.1.3. Autonomy

The Autonomy item is designed to assess the ability to be independent, both in the practical sense of being able to look after oneself and in being able to make decisions [115]. A small number of informants discussed independence as a dimension of quality of life prior to seeing the measure. The term "independence" was not used, rather discussion by these informants focused on the ability of an individual to do day-to-day things, such as shopping, which was often closely linked by informants with a person's mobility.

IS06 ...they can't get down to the shops to do their shopping. Or, they want to do something and they can't do it. It is hard, they have got to think about is it feasible to do something that they want to, based on how mobile they are.

In comparison to the limited discussion prior to viewing the measure, the majority of informants identified the Autonomy item as being of central importance to the assessment of quality of life. There was a consistent opinion that the item was more important to older people.

IS17...especially with older people that independence is hugely important to them, and that's one of the depressing things for them when they lose that independence I think.

5.3.2.1.4. Achievement

The Achievement item is a measurement of the degree to which an individual can attain their goals and move forward in life [115]. The influence upon quality of life of being able to achieve and attain personal goals was not discussed by many informants prior to viewing the measure. Gaining a sense of achievement through work was discussed briefly by a small number of informants. The importance of being able to look back with a sense of achievement was noted as important.

IS01 ...I think he [young cancer sufferer] has kind of condensed it all to "Yeah, I am 25 and I have achieved everything I want"...and he is perfectly sane in what he is saying, be he has just reflected back and gone, "Yeah, I have achieved".

Informants provided greater discussion after seeing the measure. The item was considered to be relevant, although disagreement existed as to whether it was relevant for older people. The item was considered by a number of informants as being too broad and for some this raised the question about whether the top level was really achievable.

IS03 I mean I don't think that I can achieve and progress in all aspects of my life, I would love to be able to. BUT.

The use of the word "progress" was questioned as for some informants in meant something other than what the question was assessing. For some it focused on the area of paid employment (e.g. progress of work), while some of those who worked in cancer noted that patients could misunderstand the question as assessing their illness (e.g. has the cancer progressed).

> *IS01 I think some of them [patients] might think progress as in is the treatment working. Or maybe that is my cancer background. Is there any progress, has the cancer progressed?*

5.3.2.1.5. Enjoyment

The Enjoyment item assesses the enjoyment gained from fun and exciting things, as well as the simple pleasures in life [115]. Enjoyment was discussed by a few informants from the perspective of people with illnesses or disabilities enjoying life in spite of their condition. It was normally identified through providing examples, rather than stating explicitly that enjoyment was a construct of quality of life.

> *IS04 You have people that have an enormously great quality of life who can't walk anywhere...because they have this great social structure and play cards all day.*

On considering the item, informants were split between those who felt the attribute was important and relevant, and those who did not. For those who felt the item was not relevant, a motivating factor appeared to be that it was too broad to be relevant. Field notes taken by the interviewer noted a level of surprise by some informants upon seeing the Enjoyment item.

> *IS18 What do you mean by enjoyment and pleasure? I suppose not vague, but possibly ambiguous.*
> *IS01 I think it is very important to have certain feelings...like enjoyment and pleasure.*

5.4. Summary of content and face validity results

Informants viewed quality of life as a broad construct, influenced by physical health, psychological health and a number of social factors. Informants made direct statements about the breadth of the construct, which they thought to be broader than health, and described a high and diverse number of influences. The comparison of the informants' conceptualisation of quality of life prior to viewing the measure with the descriptive system of the ICECAP-A indicates that during the discussion, in the first part of the interview, informants discussed all of the items included in the ICECAP-A descriptive system, as influences upon quality of life. However, it should be noted that there was a considerable amount of discussion about the importance of physical and psychological health and the affect that it can have on quality of life. Despite the breadth of the measure, physical and psychological health are not directly assessed by the ICECAP-A. Determinants which informants held as having a notable impact on quality of life, such as pain and side-effects, are not directly assessed by the measure.

The results indicate that on viewing the ICECAP-A informants felt that it was a useful, broad measure of quality of life, which captured many aspects of quality of life that were discussed in the first part of the interview. Informants' perceptions of the ICECAP-A were that it measured the broad construct that they had previously defined. A number of informants highlighted that it did not directly assess the health of the individual, which informants had previously defined as an important aspect of quality of life. Item-by-item analysis of the discussion in the second part of the interview showed informants emphasised the importance of the Stability, Autonomy and Attachment items; while questioning the relevance of the Enjoyment item, and to a lesser extent the Achievement item.

5.5. Selection of quality of life measures for use in randomised controlled trial.

A theme which emerged through the interviews and analysis was around the decision making process about which quality of life measures are selected for use in a trial research environment. Exploration of this issue was not an aim of the interviews at the outset. The original topic guide had no prompts or probes pertaining to this topic, but in discussing the pros and cons of existing measures in the early part of the interview, and when discussing the ICECAP-A and EQ-5D-5L, it was natural that informants would explain why measures were chosen. In the early interviews informants made reference to both how and by whom a measure was selected and what the primary considerations during this selection process were.

This emerging theme was identified as potentially important to furthering understanding of the use of quality of life measures in a trial setting. The topic guide and the analysis codes were refined in order to identify data relating to this theme. In subsequent interviews this theme then continued to be discussed through the use of scoping and mining probes. While much of the data referring to this theme was not discussed by informants when referring directly to the ICECAP-A measure, the theme is important for understanding the potential barriers to use of the ICECAP-A measure.

Data collection was stopped once saturation had been reached on the primary focus of the research and data saturation was not reached on this emerging theme. The data collected provide an initial indication of a) the factors taken into account by those deciding on quality of life measures b) who decides which quality of life measures were used in trials and what

the decision process is and c) what the most likely decision is. Data have allowed the tentative proposal of a conceptual model for the decision process of selecting a quality of life measure, which could be tested and refined through further work.

5.5.1. Factors taken into account when choosing a measure for use

Five factors were identified as important when choosing a quality of life measure for use in a research trial setting: precedent of use, evidence of validity, evidence of sensitivity to change, perceived relevance of content and practical considerations (e.g. respondent burden). An indication of the relative importance of each was assessed by the consistency with which each consideration was identified and the terms in which informants discussed it. Precedent of use and evidence of sensitivity to change were identified as potentially two of the most important considerations.

5.5.1.1. Precedent of use

Precedent of use appears to be an important factor in choosing a measure for use in a trial. Whether the measure had been previously used and was widely accepted amongst research colleagues, or had been recommended by an authoritative body, appeared to be a key motivator when deciding on the appropriate measure. This appeared to be motivated in part by a desire to be able to compare across studies and interventions.

> *IS04 You know, I haven't thought about it...is it good is it bad... no....it is what is accepted, expected. It is what is there. It is not a major focus.*
>
> *IS08 And then it is just being used because you want to say something about your results in light of other results. So it if you can't compare it to anything then that makes it hard*
>
> *IS19 But you know at the end of the day it is the best tool that we have, or it is the tool that is most often used, and there is a case consistent from one setting to another.*

For informants who had knowledge of economic analysis alongside clinical trials, the importance of precedence in the decision process had been reinforced by NICE guidance. NICE guidance was identified as the reason that the EQ-5D had gained such dominance in health economic analysis alongside clinical trials.

> *IS18 Because NICE recommends it and it is not too onerous.*
>
> *IS16 it is kind of a bit of a mantra I guess. It would be a fairly assertive person [health economist] not to have EQ-5D in.*

There was also a tendency for informants to defer their opinion to that of experts. A number of examples were found of informants assuming the measure was suitable for use in the trial because the measure had been designed and validated by quality of life "experts".

> *IS08 There's a good team behind the EORTC, the EORTC is a great organisation I think …my first thought is the EORTC...*
>
> *TK Do you think they capture quality of life adequately?*
> *[pause]*
> *IS13 Who am I to say? Experts have designed these to try and capture domains which cover various, must come close to covering [quality of life].*

5.5.1.2. Evidence of validity

A number of informants who were involved in the selection of measures stated that they would want to see evidence of validation or know that a measure had been validated before use.

> *IS10 But, if someone talks about a questionnaire I have never heard about I would want some reassurance that it has been validated and the properties are well established.*
>
> *IS02 I would want to know a questionnaire was appropriately validated.*

This shows a tension with the theme of precedent of use. Here informants want to see validity evidence rather than deferring opinion either to previous precedent of use or to experts. It

may be that through precedent of use validity evidence is gained, or precedent use is how informants judge a measure as valid. Data cannot confirm this.

5.5.1.3. Evidence of sensitivity to change

Sensitivity to change was identified as a major concern for informants. It was the most important of the psychometric properties (validity, reliability and sensitivity to change) discussed.

> *TK and that sensitivity to change is a big concern?*
> *IS16 YES [laughs] Yes huge.*

A concern existed amongst the majority of informants that quality of life measures were not sensitive to change and this made them cautious about using them in their trials.

> *IS16 I have never used a quality of life measure as a primary outcome…I have shied away from using it basically because I haven't seen any evidence beforehand that I would be able to demonstrate a change in quality of life.*

This concern was particularly evident when discussing EQ-5D-3L. Despite its frequent use by informants, the majority felt the measure was not adequately sensitive to change. This may indicate that precedent for use rather than sensitivity to change has a primacy in considerations when selecting a measure.

> *IS06 I guess in terms of the measurement [EQ-5D-3L] …it is a bit unclear whether you can actually detect affect from our intervention, so the jury is still out on that.*

5.5.1.4. Perceived relevance of content

Perceived relevance to the trial population of the content of the measure was seen as being important. Informants recognised that there are differences both in the descriptive systems of quality of life measures and in those aspects of quality of life that are relevant to different trial populations. This point was often raised when discussing the measures under consideration in these interviews. Informants identified measures as being particularly suitable and relevant,

or not, to different research populations with which they work. For example: informants often noted that the EQ-5D-5L descriptive system would suffer from a ceiling effect in healthy research populations; while the ICECAP-A may lack a health focus in severely sick populations.

> *IS07 Well its quite targeted to the population so it's not, you know you haven't got most of the people just saying there is not a problem on all these things, it has to be appropriate, the questions have to be appropriate to the target population.*
>
> *IS10 I still feel this would be better in general population*

5.5.1.5. Respondent burden

Practical considerations focused primarily on the length of the questionnaire and the time it would take to complete the measure. Informants identified a need for concise questionnaires, while at the same time recognising there was a trade-off between questionnaire length and completeness of coverage.

> *IS01 because they haven't got the time to spend with the long questionnaire and you're not going to get attention for long, because they'll be looking to see are they going to get called in to get there bloods done, are they waiting for the doctor. So it has to be quite quick, quite easy*

5.5.2. Decision process when choosing a measure for use

Informants described a collaborative process through which quality of life measures were chosen for use in a trial. Informants reported that the steering group, led by the chief investigator, generally decides on the inclusion and exclusion of measures. Research clinicians and health economists may be involved in the trial management group, but are often less involved in the running of the trial. Those involved in the decision process are referred to as agents.

> *IS18 So yeah every trial has a different group people around it the table. There is*

usually a lot of horse trading. Usually everything is bunged in and then they say this is going to take two hours
[laughs]
TK A hundred page baseline questionnaire
[laughs]
IS18 yeah quite...it is negotiated overall, the whole package of outcome measures is negotiated.

IS13 A combination of the steering committee, the chief investigators...and experts they will bring in to run the study.

Non-health economist informants, including statisticians and clinicians, identified health economists as "experts" on quality of life measures. Health economists described how they were in charge of selecting the health economics outcome measures, including preference based quality of life measures.

IS19 The things feeding into the economic analysis I would have control of.

The health economists did not however feel they were responsible for, or capable of, making decisions for all quality of life measures. Health economists did not view themselves as experts on all quality of life measures and identified research doctors as being more appropriate judges of the broader spectrum of quality of life measures, including disease specific measures.

IS19 Other quality of life instruments which are nothing to do with the health economics, I wouldn't have that, because I wouldn't be in a position of being knowledgeable in those questionnaires. It is normally the clinicians who will be up to speed on the literature surrounding that particular intervention who would know that other studies report measure X Y and Z.

On the role of research doctors a degree of agreement may be present in the data. The informants who normally discussed the relevance of the content of the measure for their populations were research doctors. They brought a level of analysis and perspective that was not present from trialists or health economists.

> *IS10 if it's an oncology population and people are really sick and people are going to die, that's probably a bit too broad...it's going to miss the point.*

However, all three research doctors noted in the early stages of the interview that quality of life was not their field of expertise and were cautious about commenting on it.

One informant described a role for an industry agent in determining what quality of life measure was to be used in industry sponsored trials. A research doctor described how industry agents wanted a measure that would show a difference for their product. It was also indicated that this decision would be made outside of a collaborative environment and might be "imposed" upon the trial.

5.5.3. Likely decisions and dominant considerations

There was some indication in the data about what the most likely decision of each of the agents in the decision process would be. Health economists seem likely to choose an accepted preference based quality of life measure, acceptable for use in health economic evaluation. Precedent of use seems to be the dominant consideration for health economists. All health economists identified concerns and perceived failings with the EQ-5D-3L (which focused around sensitivity and relevance), but equally all health economist informants said that they used the measure.

> *IS18 But I always say "the EQ-5D we have got to have it in there"...even if you don't think much of it, it has got to be in there.*

A collaborative decision taken, by the trial steering or management group, informed by doctors, health economists and statisticians, will likely choose a measure that is comprehensive, appropriate to the population and sensitive to change. Precedent of use may be considered directly or used to judge these qualities and the practical decisions of length and

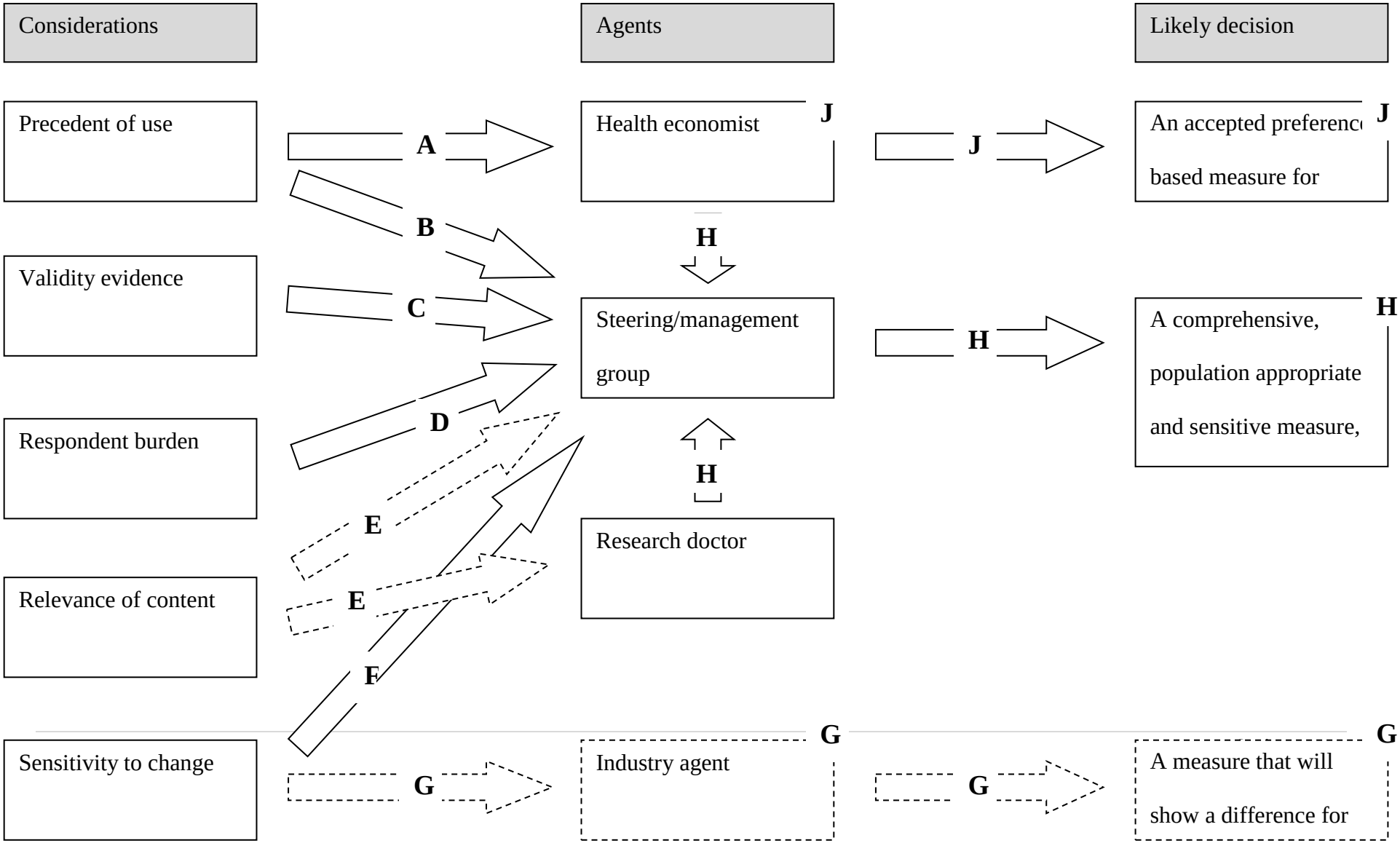
time of completion would exert some influence. The collaborative decision would likely select a non-preference based measure such as the SF-36 or a disease specific measure like the EORTC QLQ C-30.

The very limited amount of data collected on the decisions of industry indicated that a measure that would show a positive outcome would likely be selected. The primary motivation here was a measure that was sensitive to change and would provide a "selling point" for their product.

5.5.4. A conceptual model of quality of life selection

A conceptual model is proposed based on the exploratory data collected from these interviews (Figure 11). This model proposed possible considerations, how these considerations may influence the different agents involved in the selection of quality of life measures and what the likely decisions of these agents may be. This model is a tentative suggestion of the process by which a measure may be selected. The model is based on exploratory data, which (for reasons stated above) did not reach data saturation during the interviews conducted. Dotted lines indicate that the evidence collected was limited and no firm conclusion could be made.

Figure 11: A conceptual model of quality of life measure selection for use in randomised controlled trials



Explanation of Figure 11.

A - Precedent of use for the health economist is focused largely around what their colleagues have used in the field and the reference case of a funding or rationing body.

B – Precedent of use for the steering and management group appears to be motivated more by the deferring of opinion to precedent. An assumption made that because it has been used by colleagues then it "must" be appropriate.

C – The consideration of validity evidence was mentioned by a small number of informants. It appears that depending on the experience and knowledge of individuals in the steering group that there will be differing levels of consideration the state of the validity literature on the measure.

D – The practical considerations were described as "horse trading" focused on a trade off between completeness and brevity of the questionnaire pack. Data indicates that this is done collaboratively with the PI having the final say.

E – Those who commented as relevance of content being a consideration were largely medically trained, and all but one were research doctors. Research doctors do not appear to make quality of life instruments decisions outside of the collaborative setting of the steering group; however it appears that they *may* bring consideration of relevance to this discussion. It is not possible to say with certainty that research doctors provide this consideration.

F – Sensitivity to change is a major consideration of the steering group, possibly to the extent that it has a primacy among the considerations.

G – Only one informant commented on this pathway. He was very clear that the consideration of an industry agent would be focused primarily/exclusively on the ability of the measure to show change.

H – Health economists and research doctors work in a collaborative fashion with the steering and management group to select quality of life measures. This collaborative decision will likely lead to a comprehensive, population appropriate and sensitive measure, with a low patient burden being selected. These measures may be disease specific, such as the EORTC QLQ C-30 in cancer or of greater length (than health economic measures) like the SF-36 or the HUI.

J – Health economists have the determining say on health economic outcome measures and will likely select an accepted, generic, preference based measure such as the EQ-5D. This may be decided at through the collaborative nature of the steering group; however the locus of responsibility for the decision is firmly with the health economist. This is the perspective of both health economists and other members of a steering group.

5.6. Summary of measure selection results

This emergent theme indicates that there are a number of factors considered when choosing a quality of life measure for use in a randomised controlled trial. For new quality of life measures these factors may be considered barriers for use. New measures will not have precedent of use and, potentially, only early indications of validity from the development process. The process of assessing the psychometric properties of the ICECAP measures is underway and both measures have started to be used in a research setting. However, in comparison to other quality of life measures, both the precedent of use and the weight of evidence are small. Therefore, this emergent theme highlights the importance of the assessment of validity and responsiveness in allowing trials to include the ICECAP measures and the potential that there may be resistance to using these measures until a clear precedent of use exists.

CHAPTER 6. QUANTITATIVE STUDY OF THE VALIDITY AND RESPONSIVENESS OF THE ICECAP MEASURES: METHODS

6.1. Chapter introduction

This chapter reports the quantitative methods used to assess the psychometric properties of the ICECAP capability measures in a randomised controlled trial setting. Two broad lines of investigation were conducted: 1) an investigation of construct validity and 2) an investigation of the responsiveness of the measure to change over time. The reliability of the ICECAP measures will not be assessed in this thesis as a) the conditions do not exist for a test re-test analysis and b) measures of internal consistency are not suitable for the ICECAP measure (both points are discussed in greater depth in section 3.6.1. Reliability). The data for this research was collected from the BEEP trial at Keele University, in which the ICECAP-A measure was included, and the PastBP trial at the University of Birmingham, in which the ICECAP-O was included in. A description of these trials is given, and the process by which they were recruited to this research is described. The process of evidence-based hypotheses formation, the hypotheses which were formed, and the statistical methods used to test these hypotheses are described. The analyses used to assess the validity and the responsiveness of the measures are provided. The results of the construct validity analysis can be found in Chapter 7 and the results of the responsiveness to change analysis in Chapter 8.

6.2. Trial recruitment and data

Two medium sized multi-centre randomised controlled trials were used to assess the validity and responsiveness of the ICECAP-A and ICECAP-O measures. The ICECAP measures were included in the baseline and follow-up questionnaire packs which were completed by patients. Access to the data was granted by the trial teams for the purpose of assessing the psychometric properties of the ICECAP measures.

6.2.1. Recruitment of trials to quantitative research

The involvement and cooperation of research groups external to the PhD supervisory team was required to test the psychometric properties of the ICECAP measures in a randomised controlled trial setting. In November, 2009 the process of "recruiting" trials to include an ICECAP measure began. The objective at the start of the process was to recruit between two and four trials, in different clinical areas. The trial researchers had to be willing to allow external analysis of data for the purpose of validating the ICECAP measures. The primary challenge faced was to identify trials that were soon to start and would yield data within the time-span of the PhD (i.e. before summer 2013). A two-step recruitment process was completed: step one focused on formally approaching trial units directly; step two sought to use contacts with the Health Economics Unit and the MRC Hub for Trials Methodology Research at the University of Birmingham, who were working with trial groups[4].

In the first step three trial groups were approached: Cancer Research UK Trials Unit, Birmingham; Birmingham Clinical Trials Unit, Birmingham and Primary Care Trials Unit,

[4] Unlike the qualitative research, trials were only recruited from UK institutions. At the time of recruitment of trials the professional links which led to the qualitative research in Australia had not been formed.

Birmingham. Trials units were approached through an email requesting assistance (Appendix 11). If a positive response was received then a meeting with the appropriate staff members was scheduled. Positive responses were received from all units and meetings were held during December, 2009.

Productive meetings were held with all three trials units. Initially three potential trials or pilot studies were identified from these meetings. A pilot study into the effects of diet modification on bowel cancer risk in those at risk of bowel cancer was identified. Contact was kept with the trial until late 2010 when the Principal Investigator on the trial withdrew co-operation for using the ICECAP-A citing concerns that the content of the questions may prove upsetting or disturbing for participants in his trial. A second surgical trial was discussed which looked at the effectiveness of a wound guard in preventing post-operative infections. Contact was maintained with this trial until the surgeons on the investigative team withdrew support in early 2011 citing concerns about whether the ICECAP-A measure would be responsive to the changes that their intervention would bring about. A third trial, the PastBP trial (described in detail below), was recruited to the research and provided data for this analysis.

The second step of the recruitment was completed through the Health Economics Unit and the MRC Hub for Trials Methodology Research. The Health Economics Unit completes the economic evaluation alongside a number of trials. The MRC Midlands Hub for Trial Methodology Research conducts research into trial methodology and holds an advisory role with trials units. Therefore, both groups have a number of links with trials units that were of potential use to this research. Two studies were identified as potentially appropriate for inclusion of the ICECAP-A measure. The first trial, assessing the efficacy of a treatment for non-operable liver tumours, was identified in 2009. In the initial meeting researchers from

the trial voiced concern about the content of the measure, which they felt could be upsetting for people suffering from cancer. Contact was kept with this trial until they declined to include the ICECAP-A measure in early 2010. A second trial, the BEEP trial (described in detail below), was recruited to the research and provided data for this analysis.

6.2.2. Trials agreeing to participate in quantitative research

The characteristics of the two randomised controlled trials that agreed to participate are described below.

6.2.2.1. The PastBP trial

The PastBP study (ISRCTN 20962286) was a multicentre, randomised controlled trial, run by the Primary Care Trials Unit, the University of Birmingham. The full title was 'A randomised controlled trial of different blood pressure targets for people with a history of stroke or transient ischaemic attach (TIA) in primary care' [279]. The primary aim of the PastBP trial was to determine whether 'a more intensive target blood pressure for people with stroke or TIA in a pragmatic primary care setting will lead to a lower blood pressure'$_{(p.6)}$[279]. A secondary aim was to assess the impact of intensive blood pressure monitoring on quality of life.

GP practices were recruited by the trial through the Primary Care Research Networks and the Midlands Research Practices Consortium. Participants were identified through general practices, using the TIA/Stroke register. Patients with a validated history of Stroke or TIA, a systolic blood pressure over 125mmHg and who were not taking three or more anti-hypertensives were eligible for inclusion. The intervention arm consisted of a target systolic blood pressure of 130 mmHg, or a 10 mmHg reduction in systolic blood pressure (BP) if the

subjects BP is below 140 at baseline. The control arm had a target of 140 mmHg systolic blood pressure. The intervention lasted for 12 months.

The primary outcome of this trial was change in systolic blood pressure between baseline and 12 month follow up. Secondary outcomes included quality of life and adverse events. Follow-up was at 6 and 12 months, with intermediate follow-ups at monthly intervals between 1 and 3 months. The ICECAP-O measure was completed at baseline, 6 and 12 month follow-ups. The 12 month follow-up data were used for the analysis of responsiveness in this thesis.

A target of 610 patients (305 in each arm) gave the trial a 90% power at a 5% confidence level to detect a 5mmHg difference in systolic blood pressure. 1167 patients attended baseline. 529 were randomised. The low randomisation rate was due to people not meeting the randomisation criteria. Specifically, the number of people who had lower than the required blood pressure was higher than expected. Further details of the PastBP trial participants are provided at the start of both quantitative results chapters.

The PastBP trial had a number of advantages for assessing the psychometric properties of the ICECAP-O measure. First, due to the inclusion/exclusion criteria, the older people recruited by the trial were suitable for completing the ICECAP-O. Second, this population also had a history of stroke or TIA, which indicated that they would be a less healthy sample in comparison to the general population. The development of the ICECAP-O was completed with samples of the general population. Therefore, the use of a clinical population offers the potential to establish validity in other contexts and for the triangulation of data. The third advantage was that this trial had a short follow-up period and was very likely to provide data with the time-frame of this doctoral research. Finally, a large number of health and health-

related quality of life measures were completed by participants, providing ample opportunity to assess validity and responsiveness.

6.2.2.2. The BEEP trial

The BEEP trial (ISRCTN 93634563) was a multi-centre, randomised, pragmatic controlled trial to assess the clinical and cost-effectiveness of, and adherence to, individually tailored, physiotherapy led exercise interventions compared to normal physiotherapy care [280]. The full title of the trial was 'Improving the effectiveness of exercise for knee pain in older adults in primary care: Benefits of Effective Exercise for knee Pain (BEEP)' [280]. The trial was part of a collaborative link between Keele University and the University of Birmingham. Potential participants were identified through either a general practice record search, a survey of older adults registered with participating practices, or from a list of patients currently being referred to physiotherapy for knee pain.

Participants were randomised to one of three groups: a control group of standard physiotherapy care which lasted 12 weeks; an intervention group of individually tailored exercise that lasted 12 weeks, including more contact sessions than standard care; and a target exercise adherence arm, which lasted up to 6 months [280]. Follow-up was via self-complete postal questionnaires completed at 3, 6, 9, 18, 36 months. The trial randomised 526 participants at baseline. The 6 month follow-up data were used in the analysis of responsiveness. This was the only follow-up data available within the time-frame of the thesis.

Due to an administrative error, an early version of the ICECAP-A was mistakenly included in the baseline questionnaire packs for 70 participants. The remaining 456 questionnaire packs

included the finalised version of the ICECAP-A. The study ID's of the 70 participants were identified through a manual check of paper questionnaire packs by TK. All follow-up questionnaire packs received the finalised version of the ICECAP-A. The difference between the early version and the final version was slight, with the word "only" being included in one response option for each item in the early version. Despite descriptive analyses and simple statistical checks finding little or no difference between the response patterns for the two groups, the patients receiving the early version of the ICECAP-A were excluded from the analysis. This cautious approach was followed based on the reasoning that when assessing the psychometric properties of a test, use of the final version is an important pre-requisite.

The primary outcome measure for the BEEP trial was self-reported pain and functioning on the Western Ontario and McMaster Universities Arthritis Index (WOMAC) questionnaire [281]. A number of secondary outcome measures were used including measures of pain, functioning, depression , anxiety and quality of life.

This study has a number of advantages for assessing the psychometric properties of the ICECAP-A measure. First, the BEEP trial population includes adults across the age spectrum and is therefore suitable for the ICECAP-A measure. Second, the majority of the participants in BEEP have a joint problem, often due to arthritis; this is a different disease pathology to Stroke and TIA seen in the PastBP trial. Third, a large number of health and health-related quality of life measures were completed by informants, including assessments of psychical and psychological health.

6.2.2.3. Measures included in trials

A number of self-report measures were included in the BEEP and PastBP trials, the data from which were provided for this validity and responsiveness analysis. The measures are summarised in Table 6.

Table 6: Self-report outcome measures included in BEEP and PastBP trials and used in quantitative analyses

Construct	Measure	Description	Trial containing measure
General health	EQ-5D-3L	The EQ-5D-3L is a generic preference based outcome measure, which measures health-related quality of life [142,143,282,283] (Appendix 12). The descriptive system comprises: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression [284], with three response options in each dimension. It is scored via a preference weighted algorithm, which for UK values produces a score between -0.59 and 1. The EQ-5D-3L has been extensively validated in numerous clinical settings [285–287].	BEEP PastBP
General health	SF-36	The SF-36 is a survey of patient health status consisting of eight sub-scales [144] (Appendix 13). The sub-scales are: physical functioning, physical role, bodily pain, general health, vitality, social functioning, emotional role and mental health. Each scale is calculated from a number of questions within the questionnaire which are given equal weight during the transformation of the scale onto a 0 to 100 scale. These 0 to 100 scales can be transformed into a T-scores, in line with normal practice for scoring of the SF-36 [288]. A T-score has the advantage of standardising scores against a population norm which is fixed at 50. This allows easier interpretation of scale scores (e.g. below 50 on any scale is a worse health state than the population norm) and has the advantage that published minimally important differences can be used. The disadvantage of using T-scores is that US population norms have to be used [289]. The measure has been extensively validated in a number of different clinical and administrative settings [286,290]. It is possible to calculate SF-6D from this measure. Although this wasn't done here as it would have resulted in a loss of information provided by the sub-scales	PastBP

Symptoms/ side-effects and co-morbidities	Non-validated measures used	Symptoms and side-effects or co-morbidity data was collected by both trials. The BEEP trial asked a total of 10 questions on co-morbidities (Appendix 14). In the PastBP trial a total of 24 questions were asked about symptoms and side effects (Appendix 15). Differences other than the number of questions asked existed between the two questionnaires. The Past BP asked it the participant had experienced symptoms, while the BEEP trial asked if the participant had ever been told they had any of the following. The PastBP trial included minor side-effects such as sore throat, while the BEEP trial focused on serious comorbidities such as heart attack, diabetes and depression.	BEEP PastBP
Mobility and physical functioning	WOMAC	The WOMAC is a widely used set of standardised questions suitable for evaluating the condition of patients with osteoarthritis of the knee and hip (Appendix 16) [281,291]. The measure has 3 sub-scales: pain, assessed by 5 questions; stiffness, assessed by 2 questions; and physical functioning assessed by 17 questions. The range of values for each sub-scale is: pain 0 to 20, stiffness 0 to 8, functioning 0 to 62. All questions carry equal weight in the calculation of sub-scale values. The measure has been validated against other measures of health and for use in different clinical and administrative settings [281,291–293].	BEEP
Illness perception	Brief IPQ	The Brief Illness Perception Questionnaire (IPQ) assesses patients' ideas and perceptions of their illnesses or conditions [294]. The questionnaire is designed so it can be adapted so as to ask about the ideas and perception of a specific illness or condition, in this case knee pain, through a simple re-wording of the questionnaire (Appendix 17). For example, the question "How much does your knee pain effect your life, could be tailored to ask about cancer simply by replacing "knee pain" with "cancer" at the appropriate point. There is some, limited, evidence of validity and reliability for the Brief IPQ [294].	BEEP

Psychological health	GAD-7 PHQ-8	The Generalised Anxiety Disorder Assessment (GAD-7) and the Patient Health Questionnaire depression scale (PHQ-8) (Appendix 18) are two short questionnaires that assess anxiety [295] and depression [296] respectively, which have validity portfolios [296,297]. They are frequently used in clinical practice. The value range is from 0 to 21 for the GAD-7 and 0 to 24 for the PHQ-8. Both measures have clinically meaningful cut-offs of 8 for the GAD7 indicating possible anxiety disorder and 10 for the PHQ-8 indicating possible depression.	BEEP
Disability	Modified Rankin Scale	The Modified Rankin Scale is a scale for measuring disability or dependence in the activities of daily living in people that have suffered neurological impairment [298]. The scales ask the participant to rank themselves on a 0 to 5 scale describing the physical situation. The psychometric properties of the measure have been assessed [299].	PastBP

6.3. Assessing the construct validity of the ICECAP measures

Validation is an on-going process of building the "validity portfolio" of a measure through the use of rigorous and scientifically sound methodology [200]. Current best practice states that this should be completed through developing evidence-based hypotheses and subjecting these to extended testing and analysis [205,212].

6.3.1. Hypothesis formation

A rigorous method of evidence-based hypothesis formation was followed. Data from the review of development and validation studies of the ICECAP-A and ICECAP-O reported in Chapter 2 were used to assist hypothesis formation. These data consisted mainly of quantitative evidence, which related both directly and indirectly to the comparator measures in the BEEP and PastBP studies. For example: previous studies have used the EQ-5D-3L as a comparator in validation studies, and the EQ-5D-3L measure was included in both the BEEP and PastBP trial. In this situation data could be used directly to inform hypothesis formation. Previous studies have also used measures, such as the Short Physical Performance Battery. This measure was not included in either the BEEP trial or the PastBP trial, but is closely related to measures, such as the Modified Rankin Scale. In this situation data could be used indirectly to inform hypothesis formation.

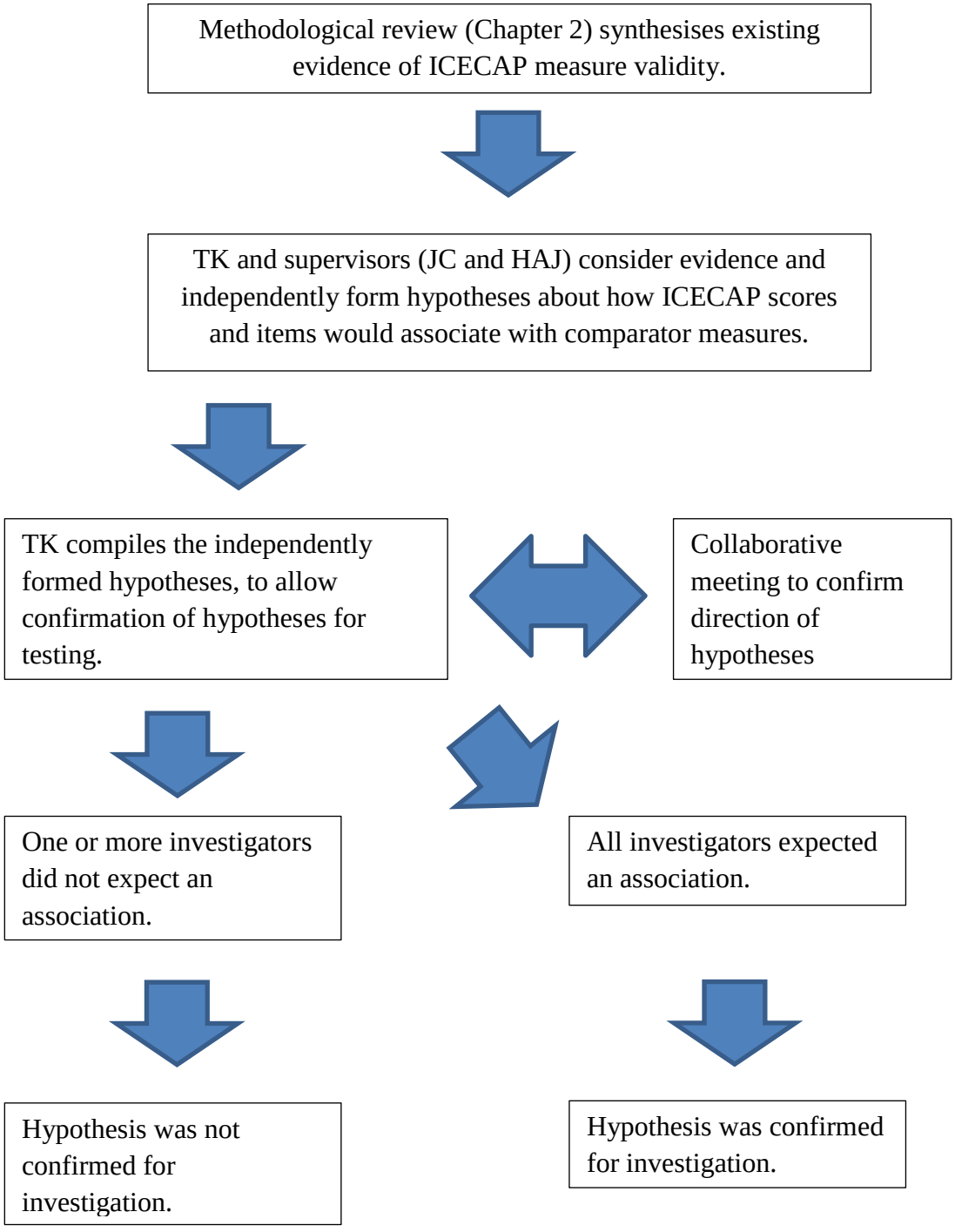
The collated information from the review was compiled by TK and presented to three investigators (TK,JC,HAJ), along with a list of the comparators in each of the trials being used in the research. Investigators independently formed hypotheses of the associations

between a comparator measure (i.e. the EQ-5D-3L, SF-36 etc), as well as sub-sections or items of the comparator, and the ICECAP-A or ICECAP-O tariff score, as well as individual item scores of the ICECAP measures. Each indicated a) whether they thought an association would be found, b) the direction of the association or correlation and c) for the EQ-5D-3L measure, which associations and relationships would be the strongest. For example: all ICECAP-A items may be expected to associate with the mobility item on the EQ-5D-3L, but the association with the ICECAP-A item of Autonomy with the mobility item might be expected to be the strongest.

The individual hypotheses were then collated through a predefined process (see Figure 12). When all three investigators agreed that an association between the ICECAP tariff or item and the comparator measure was expected, this hypothesis was confirmed for testing. If any investigator did not expect an association, the hypothesis was not confirmed for testing. In the situation where investigators expected associations in different directions, this was discussed collaboratively (it was normally due to a misunderstanding of the direction of the scoring of the comparator measures).

The three people who participated in the hypothesis formation had an in depth knowledge of the capability approach, development of the ICECAP measures and validation efforts to date. The hypotheses presented below were therefore formed based both on the theoretical knowledge of the investigators and the results from the methodological review.

Figure 12: The process of hypothesis formation



6.3.1.1. Results of the hypothesis formation

There was a high level of agreement amongst the three investigators forming hypotheses. For the ICECAP-A analysis, using BEEP trial data, investigators formed 8 hypotheses on how the ICECAP-A value tariff would associate with other measures and 70 item-by-item hypotheses. Full agreement was found between the three investigators on the 8 ICECAP-A value tariff scores. For the item-by-item hypotheses 10 disagreements were found and these hypotheses were not confirmed for testing. Therefore the overall agreement rate was 87%. For the ICECAP-O analysis, using PastBP trial data, researchers formed 12 hypotheses about how the ICECAP-O value tariff would associate with other measures and 101 item-by-item hypotheses. Full agreement was found on the ICECAP-O tariff values scores. 18 disagreements were found for the item-by-item analysis and these hypotheses were not confirmed for testing. Therefore the overall agreement rate was 84%.

6.3.2. Hypotheses

The *a priori* hypotheses for each trial are reported below. Hypotheses for the overall ICECAP score and individual items of the ICECAP measure are described. Appendix 19 and Appendix 20 show the detailed breakdown of these hypothesised associations including the expected direction of the association (taking account of the coding of the items and tariffs) and, for the EQ-5D-3L measure in both studies, which item would show the strongest association with each item of the ICECAP-A measure.

6.3.2.1. BEEP trial

Table 7 shows the hypothesised associations between the ICECAP-A value tariff and items and the physical, psychological and socio-demographic comparator measures. For each

ICECAP-A item, the phrasing of the highest response option of the item and a brief summary of the theoretical framework that surrounds this item, taken from Al-Janabi et al [115], is presented. The hypothesised associations between both the overall ICECAP-A score and the individual items of the ICECAP-A measure and measures of physical health, psychological health and socio-demographic variables are given.

Table 7: Hypothesised associations between ICECAP-A value tariff and item scores and comparator measures

ICECAP-A attribute and description from Al-Janabi et al (2012)	Physical comparators	Psychological comparators	Socio-demographic comparators
Overall ICECAP-A tariff	It was hypothesised that the overall ICECAP-A tariff score would be associated with measures of physical health. Higher levels of pain (EQ-5D-3L and WOMAC) were expected to associate with lower tariff scores. High levels of mobility (EQ-5D-3L) and physical functioning (WOMAC) were anticipated to be linked with higher ICECAP-A tariff scores. General health (EQ-5D-3L overall score) was expected to be positively associated with the tariff score. The ability to care for oneself and perform ones usual activities (EQ-5D-3L) was expected to associate with the tariff score.	It was expected that the overall ICECAP-A tariff score would be associated with measures of psychological health. High levels of anxiety and depression (EQ-5D-3L), anxiety (GAD-7) and depression (PHQ-8) were anticipated to be associated with low levels of overall capability.	It was expected that the overall ICECAP-A score would be associated with increasing age. Elderly individuals were anticipated to have lower levels of capability. No association was expected between gender and the ICECAP-A tariff score.

ICECAP-A attribute and description from Al-Janabi et al (2012)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Stability** attribute, phrased "I am able to feel settled and secure in all areas of my life", refers to the ability to live without stress, feeling threatened or dramatic changes in one's life and the ability to have continuity and attach meaning to life. This capability attribute was thought to be affected by health, unemployment, consistent relations with friends and family and secure work and finances [115].	It was hypothesised that measures of physical health would be associated with the Stability item. Poor physical health was expected to make people feel less secure and stable due the worry and concern that it causes. This may occur through making people less able to work and maintain a regular and secure wage. It was anticipated that having greater mobility and ability to self-care for oneself (EQ-5D-3L) and higher levels of physical functioning (WOMAC) would be associated with higher levels of Stability. High levels of pain (EQ-5D-3L and WOMAC), stiffness and functioning (WOMAC) were anticipated to associate with lower levels of Stability.	It was predicted that measures of psychological health would be associated with the Stability item. Feelings of anxiety, worry and depression were thought to reduce a person's ability to feel settled and secure. It was expected that higher levels of anxiety and depression (EQ-5D-3L), anxiety (GAD-7) and depression (PHQ-8) would be associated with lower levels of Stability.	Stability was not expected to be associated with gender or age.

ICECAP-A attribute and description from Al-Janabi et al (2012)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Attachment** attribute, phrased as "I can have a lot of love, friendship and support", refers to the capability to be close to people, experience affection and have a sense of belonging. Sources of these feelings include ability to interact with friends, family and partners, as well as the quality of these interactions. It is therefore closely linked to the presence of family, friends and partners and events that bring family closer together [115].	It was not hypothesised that measures of physical health would be associated with the Attachment item. Evidence from the ICECAP developmental work indicated that poor physical health of an individual might strengthen relationships and draw people closer together, while the initial validation work showed mixed results under the Attachment item.	It was not hypothesised that Attachment would associate with measures of psychological health. There was considerable lack of agreement in the hypotheses researchers formed. This disagreement was likely due conflicting findings in the previous literature highlighted by the review and presented to researchers.	Attachment was not expected to be associated with gender or age.

ICECAP-A attribute and description from Al-Janabi et al (2012)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Autonomy** attribute, phrased "I am able to be completely independent", refers to the ability to be independent, look after oneself and not feel a liability to others. It is closely linked to poor health which might impact a person's ability to complete basic activities and corrode their sense of identity [115].	It was hypothesised that Autonomy would be closely associated with measures of physical health. Good physical health is an important pre-requisite and determinant of independence. It was anticipated that greater mobility and self-care (EQ-5D-3L) and physical functioning (WOMAC) would be associated with higher levels of Autonomy. It was expected that these measures of physical health and functioning would show a stronger association with Autonomy than other ICECAP-A items. High levels of pain (EQ-5D-3L and WOMAC) and stiffness (WOMAC) were expected to associate with lower levels of Autonomy. The ability of a person to complete their usual activities, a measure of independence in the activities of daily living, was thought to assess a similar domain of quality of life to the Autonomy item. Therefore, greater ability to complete usual activities (EQ-5D-3L) was expected to associate strongly with Autonomy.	It was hypothesised that measures of psychological health would be associated with Autonomy. Poor psychological health may reduce an individual's ability to look after themselves and make decisions. High levels of anxiety and depression (EQ-5D-3L, GAD-7 and PHQ-8) were expected to associate with lower levels of Autonomy.	It was hypothesised that increased age would be associated with independence. Therefore the socio-demographic variable of age was expected to associate with Autonomy.

ICECAP-A attribute and description from Al-Janabi et al (2012)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Achievement** attribute, phrased "I can achieve and progress in all aspects of my life", refers to the ability of people to move forward in order to achieve goals in their lives, as well as to look back with a sense of satisfaction on their achievements. It is closely linked to opportunities that exist at work and participation in voluntary activities and sport [115].	It was expected that the Achievement item would be associated with measures of physical health. Poor physical health was expected to reduce the activities that a person is able to participate in (i.e. work, voluntary roles, child care). Higher levels of mobility and self-care (EQ-5D-3L) and physical functioning (WOMAC) were anticipated to be associated with higher levels of Achievement capability. Lower levels of Achievement capability were anticipated to associate with higher levels of pain (EQ-5D-3L and WOMAC) and stiffness (WOMAC). The ability of a person to complete their usual activities (EQ-5D-3L), a measure of independence in the activities of daily living, was expected to associate strongly with Achievement.	It was hypothesised that measures of psychological health would be associated with Achievement. Poor psychological health, such as feelings of anxiety and depression, may reduce a person's ability to achieve and progress in their life. High levels of anxiety and depression (EQ-5D-3L) were expected to associate with lower levels of Achievement[5].	Achievement was not expected to be associated with age or gender.

[5] Due to what appeared to be inconsistency in a researchers hypothesis formation, GAD-7 and PHQ-8 was not confirmed for testing due to disagreement, whereas the EQ-5D-3L item of anxiety and depression was confirmed for testing.

ICECAP-A attribute and description from Al-Janabi et al (2012)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Enjoyment** attribute, phrased "I can have a lot of enjoyment and pleasure", refers to the ability of people to enjoy pleasurable and fun activities and live free of depression and pain. The capability for enjoyment was dependent upon a range of influences from friends and family to financial activities and poor health [115].	It was hypothesised that the Enjoyment item would be associated with measures of physical health. Good health may increase enjoyment through the enabling of enjoyable activities and increasing the quality of those activities. Higher levels of mobility and self-care (EQ-5D-3L) and physical functioning were predicted to associate with higher levels of Enjoyment capability. High levels of pain (EQ-5D-3L and WOMAC) and stiffness (WOMAC) were expected to associate with lower levels of Enjoyment and this association was expected to be stronger than for other ICECAP-A items. The activities and action of life should be a source of enjoyment. Therefore, the ability of a person to complete their usual activities (EQ-5D-3L), a measure of independence in the activities of daily living, was expected to associate with Enjoyment.	It was hypothesised that measures of psychological health would be associated with Enjoyment. Anxiety and depression was expected to reduce the enjoyment someone finds in their life. High levels of anxiety and depression (EQ-5D-3L, GAD-7 and PHQ-8) were anticipated to be associated with low levels of Enjoyment. It was anticipated that this association of psychological health with Enjoyment would be stronger than for other ICECAP-A items.	Enjoyment was not expected to be associated with age or gender.

6.3.2.2. PastBP trial

Table 8 shows the hypothesised associations between the ICECAP-O value tariff and items and the comparator measures. For each ICECAP-O item, the phrasing of the highest response option of the item and a brief summary of the theoretical framework that surrounds this item, taken from Grewal et al [109], is presented. The hypothesised associations between both the overall ICECAP-O score and the individual items of the ICECAP-O measure and measures of physical health, psychological health and socio-demographic variables are given.

Table 8: Hypothesised associations between ICECAP-O value tariff and item scores and comparator measures

ICECAP-A attribute and description from Grewal et al (2006)	Physical comparators	Psychological comparators	Socio-demographic comparators
Overall ICECAP-O tariff	It was hypothesised that the ICECAP-O tariff score would be associated with measures of physical health. Higher levels of pain (EQ-5D-3L and SF-36) were expected to associate with lower capability scores. High levels of mobility (EQ-5D-3L) and physical functioning (SF-36) were anticipated to be linked with higher levels of overall capability. High levels of disability (MRS) were expected to associate with low ICECAP-O scores. General health (EQ-5D-3L overall score and SF-36) was expected to be positively associated with overall capability. The ability to care for oneself and perform ones usual activities (EQ-5D-3L) was expected to associate with the overall capability score.	It was expected that the overall ICECAP-O score would be associated with measure of psychological health. High levels of anxiety and depression (EQ-5D-3L) was anticipated to be associated with low levels of overall capability. Higher levels of mental health, emotional functioning and vitality (SF-36) were expected to be associated with higher ICECAP-O tariff scores.	It was anticipated that the overall ICECAP-O score would be associated with the socio-demographic variable of age, but not gender.

ICECAP-A attribute and description from Grewal et al (2006)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Attachment** attribute, phrased "I can have all the love and friendship that I want", refers to the ability of people to experience of love, friendship and affection and have a sense of companionship. The sources of this are thought to be partners, family, friends and in some cases, owning pets [109].	It was not hypothesised that the Attachment item would be associated with measures of physical health. Feelings of love and affection were not thought to be dependent upon health. Evidence from the Attachment item in the ICECAP-A offers limited evidence that poor health may pull a family and friends closer together.	It was not hypothesised that measures of psychological health would associate with the Attachment item. There was notable disagreement between investigators during the hypothesis formation in this area, possibly generated by inconsistencies in previous data, with some studies showing associations and other not. However social function and emotional role, which are scales in the psychological summary section of the SF-36, were expected to associate with Attachment.	Attachment was not expected to be associated with age or gender.

ICECAP-A attribute and description from Grewal et al (2006)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Security** attribute, phrased "I can think about the future without any concern", refers to the ability to feel safe and secure, living without worry and not having feelings of vulnerability. It is thought to be affected by having sufficient finances, practical and emotional support and good health [109].	It was hypothesised that measures of physical health would associate with the Security item. Physical health problems are likely to cause greater worry and concern and reduce the feelings of security that a person is able to experience. It was anticipated that high levels of self-care (EQ-5D-3L), physical functioning, physical role and general health (SF-36) would be associated with higher levels of Security. Being in physical pain (EQ-5D-3L and SF-36) or reporting high levels of disability (MRS) was expected decrease feelings of security.	It was anticipated that measures of psychological health would associate with the Security item. Feelings of anxiety, worry and depression were expected to reduce a person's ability to think about the future without concern. High levels of anxiety and depression (EQ-5D-3L) and low levels of mental health (SF-36) were expected to associate with lower levels of Security.	Security was not expected to be associated with age or gender.

ICECAP-A attribute and description from Grewal et al (2006)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Role** attribute, phrased "I am able to do all the things that make me feel valued", refers to the ability an individual to feel valued by others, as well as themselves. This attribute is closely linked to having a purpose or doing something worthwhile [109].	It was hypothesised that measures of physical health would associate with the Role item. Poor physical health can restrict the ability of people to fulfil roles or achieve activities which make them feel valued. Lower levels of mobility, self-care (EQ-5D-3L), physical functioning, role physical and general health (SF-36) were expected to associate with lower scores on the Role attribute. High levels of pain (EQ-5D-3L and SF-36) or high levels of disability (MRS) were expected to associate with lower scores on the Role item. The usual activities item (EQ-5D-3L) is an assessment of the ability to complete everyday activities, including work and family activities. It was expected that this item would associate strongly with the ability to do things that make you feel valued. Therefore, high scores on the usual activities item were hypothesised to associate strongly with high scores on the Role item.	It was hypothesised that measures of psychological health would be associated with the Role item. Low levels of anxiety and depression (EQ-5D-3L) and high levels of vitality, social functioning, role emotional and mental health (SF-36) we expected to associate with high levels of Role.	Role was not expected to be associated with age or gender.

ICECAP-A attribute and description from Grewal et al (2006)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Enjoyment** attribute, phrased "I can have all of the enjoyment and pleasure that I want", refers to the ability to experience pleasure and joy and a sense of satisfaction. The sources of these feelings are expected to be both personal and communal activities [109].	It was anticipated that measures of physical health would associate with the Enjoyment item. Good health may increase enjoyment through the enabling of enjoyable activities and increasing the quality of those activities. While side effects of poor health, such as pain and fatigue, may reduce the enjoyment a person has in their life. High levels of mobility, self-care (EQ-5D-3L), physical functioning, role physical and general health (SF-36) were expected to associate with high levels of Enjoyment. High levels of pain (EQ-5D-3L and SF-36) were anticipated to associate with low levels of Enjoyment and this association was expected to be stronger than for other ICECAP-A items.	It was hypothesised that measures of psychological health would associate with Enjoyment. Feelings of anxiety and depression would likely reduce the level of enjoyment in an individual's life. High levels of anxiety and depression (EQ-5D-3L) were expected to associate with low levels of Enjoyment, while high levels of vitality, social functioning, role emotional and mental health (SF-36) were expected to associate with high levels of Enjoyment. These associations with Enjoyment were expected to be stronger than for other ICECAP-O items.	Enjoyment was not expected to be associated with age or gender

ICECAP-A attribute and description from Grewal et al (2006)	Physical comparators	Psychological comparators	Socio-demographic comparators
The **Control** attribute, phrased "I am able to be completely independent", refers to the ability of a person to be independent and make their own decisions. Influences upon this attribute are thought to be psychological and physical health, as well as having sufficient means and finances so as not to rely on others [109].	It was anticipated that measures of physical health would associate with the Control item. Poor physical health may increase the need for people to rely on others and reduce their ability to complete tasks independently. High levels of mobility, self-care (EQ-5D-3L), physical functioning, role physical and general health (SF-36) were expected to associated with high levels of Control. High levels of pain (EQ-5D-3L and SF-36) and high levels of disability (MRS) were anticipated to associate with low levels of Control. It was hypothesised that measures of physical functioning would show a stronger association with control than with other items in the ICECAP-O. The ability to complete the normal actions of everyday living is a large part of independence. The usual activities (EQ-5D-3L) item was expected to associate strongly with Control.	It was hypothesised that measures of psychological health would be associated with control. Feelings of anxiety, depression or emotional instability would like reduce the control an individual feels that they can exercise upon their life. High levels of anxiety and depression (EQ-5D-3L) and low levels of vitality, social functioning, role emotional and mental health (SF-36) were expected to associate with low levels of Control.	It was hypothesised that older individuals would have lower levels of independence. Therefore, increasing age was expected to associate negatively with levels of Control

6.3.3. Statistical analyses used in validity analysis

6.3.3.1. Descriptive statistics

The mean values of the measures included in the baseline questionnaire pack were used to describe the characteristics of the sample used in the analysis. The feasibility of the measure for use was assessed using the completion rates of measures and the item-by-item missing values. An assessment of potential ceiling effects was completed using the distributions of responses (the response profile) across each of the levels of the ICECAP-A. These were considered alongside values from the general population for comparison. The full assessment of the ceiling effect, described in Section 3.5.5. Floor and Ceiling Effect, was not possible due to the lack of other capability measures in these trials that could act as comparators.

6.3.3.2. Correlation coefficient

Correlation coefficients can be used to assess the strength of the association between two continuous variables. In this thesis they were used to assess the strength of the association between the ICECAP tariff scores and other measures with an outcome as a continuous variable. Pearson's correlation coefficient was used for data that were normally distributed and did not show skew [300]. Spearman's rank correlation coefficient was used for data that showed deviation from normality [301]. Correlation values less than 0.3 were described as weak, values from 0.3 to 0.6 were described as moderate and values of 0.6 and above were described as strong. Goodman and Kruskal's Gamma is a measure of rank correlation [302]. It provides an indication of the strength and direction of association when both variables are ordinal. It provides values ranging from +1 to -1, with -1 being a perfect negative correlation.

6.3.3.3. ANOVA

One-way analysis of variance (one way ANOVA) can be used to compare the mean of a continuous variable across three or more groups [303]. One way ANOVA was used to assess the mean tariff scores of the ICECAP measures across groups formed by using the levels of items in other measures. For example, the relationship between the means of the ICECAP-A tariff value across the three levels of the mobility item of the EQ-5D-3L. Where the continuous variable was not normally distributed, the Kruskal-Wallis analysis of variance was used. To assess differences in ICECAP-A and ICECAP-O tariff scores between consecutive levels of EQ-5D-3L items, Kruskal-Wallis multiple groups comparison was used.

6.3.3.4. Chi-square

Chi-square tests allows the assessment of the presence of association between categorical variables [303]. In this thesis they were used to assess associations between the ICECAP items and the items of another measure. Where appropriate, and when computationally feasible, the Fisher's exact test was used.

6.3.3.5. Factor analysis

Factor analysis is a statistical test based on the premise that a battery of questions can be described based on a smaller number of underlying factors [184]. Factor analysis describes variability amongst a number of variables or items through the use of a smaller number of unobserved variables, known as factors [138]. If a scale is uni-dimensional then one factor should explain the variance accurately [138]. Factor analysis can also be used to test the assumption that a pool of items assesses different underlying factors. For example, Davis et al [173] used factor analysis to determine whether the five constructs of the EQ-5D-3L and the

five constructs of the ICECAP-O measure assessed different underlying factors. Factor analysis was used to further investigate the association between ICECAP items and items on the EQ-5D-3L.

Factor analysis assumes that variables are continuous and follow a normal distribution. When using categorical variables factor analysis can be performed using a polychoric distribution, which was used in this thesis. The number of factors retained was chosen with reference to the Kaiser Criterion [304], which advocates retaining factors with Eigen Values greater than 1 and using the scree plot to assess the suitability of this choice. An oblique Promax rotation [305] was used, which allows for the potential that factors are correlated [138]. Correlations between factors equal to or greater than 0.32 is considered the point at which oblique rotations are appropriate [306].

6.4. Assessing the responsiveness of the ICECAP measures

6.4.1 A methodological note

As discussed in greater depth in Section 3.6 a number of methodological challenges exist when seeking to assess the responsiveness of a capability measure in a trial setting. These challenges mean that this responsiveness analysis is designed to assess changes in scores of the ICECAP measures when a change in health occurs within a randomised controlled trial; not how they respond to changes in capability. While some assumptions can be made that a change in health may have resulted in a change in capability, no firm conclusion can be drawn. Efforts were made to assess the change in ICECAP scores when a minimally important change in health occurred, but this does not allow assessment of a minimally important difference in ICECAP measure scores. Unlike the validity analysis, which drew upon previous cross-sectional studies, there is no precedent of testing the responsiveness of an ICECAP measure using longitudinal data. Due to its novel nature, the analysis is exploratory, with a focus on understanding how the ICECAP measures changed with measures of health. With this aim, numerous anchors (more than may have been used had previous responsiveness data been available to guide and inform the analysis) have been used to provide a greater breadth of results, which might inform future analyses.

6.4.2. Anchor-based analysis

Baseline and follow-up data from the PastBP trial and BEEP trial were used to explore responsiveness. Participants who had completed both baseline and follow-up assessments were included in the analysis. An anchor based analysis was completed using measures from

the trials which were administered at baseline and follow-up. Appropriate statistical analyses were completed (below). An item by item descriptive analysis of the changes in response profile was completed.

6.4.2.1. Anchor selection

An anchor based approach uses anchor measures to assign participants to groups reflecting some degree of change between baseline and follow-up in the anchor measure [261]. In line with recommendations multiple anchors were used to assess the responsiveness of the ICECAP measures [251,260,261]. Revicki [260,261] indicates that when choosing anchors, as a rule of thumb, change correlations of 0.3 are required between the change scores of the anchor and the measure under consideration, although alternative (lower) correlation thresholds may be acceptable in some situations. In line with this advice, anchors were chosen using four considerations: a) cross-sectional correlations at baseline and follow-up between the measures, b) the change correlation between the measures, c) theoretical or methodological reasons for using the anchor and d) expectation that the analysis using the anchor might increase the understanding of the ICECAP measures.

The choice of anchor is an integral part of the responsiveness analysis and is therefore reported in detail as part of the responsiveness results. The analysis has to be started for the anchors to be chosen and it is not appropriate to report this in the methods. Justification and explanation for the choice of individual anchors is provided in the results.

6.4.2.2. Anchor group formation

For each chosen anchor, groups that have improved or worsened were defined. Three methods were used to define these groups. First, some measures produce easily defined

groups which can be used. For example, the Modified Rankin Scale is a single item measure with 5 response options. Each response option can be used to provide 5 groups, which have changed by varying degrees for analysis.

Second, published minimally important differences (MID) in the anchor were used to define groups that had changed (increased or decreased) by equal to or greater than the minimally important difference. These were groups of people who have experienced change equal to or greater than MID (mean change in these groups were greater than the MID and this mean change is presented in the analysis).

Third, when MID values were not available in the literature, and a naturally occurring group was not present, the inter-quartile range was used to define groups. The inter-quartile range is the values between which the middle 50% of a distribution lies. When groups were formed using the inter-quartile range, the change groups represented the upper and lower 25% of the distribution. When it was not appropriate to use the inter-quartile range, and a minimally important difference value was not available, an arbitrary value was selected. No assumptions should be made about the importance of this change.

The method by which the anchors groups were formed is reported in the methods. The value used, the academic source which provided this value is reported and the subsequent outcome of this group's formation is reported for each anchor.

6.4.3. Methods for assessing responsiveness

Descriptions of BEEP and PastBP participants used are provided at the start of each trial analysis. The sample of participants used in this analysis is taken from the same pool of participants that was used to assess cross-sectional validity presented in Chapter 7. No

selection of participants was made and all complete data available was used in this analysis. Participant drop out and loss to follow-up meant that sample characteristics may differ from the sample used in the validity analysis in Chapter 7 and between baseline and follow-up.

Participants included in analyses in this chapter had completed the ICECAP measure and comparator measures at baseline and follow-up. Participants who failed to complete a measure were excluded from the analysis using that anchor measure, but may have been included in analyses using other anchors if they completed those measures. Differences in the rate of measure completion meant that numbers in each individual analysis varied.

For each anchor included in the analysis of responsiveness, three analyses were completed: an item-by-item analysis, an analysis using the non-value weighted scores of the ICECAP measures and an analysis using the value weighted ICECAP-O tariff score.

6.4.3.1. Item-by-item analysis

It is likely that changes in different aspects of health and physical functioning will have a differing effect upon the items of the ICECAP measure. A response profile was used to assess this change. The response profile is the spread of responses across each level of each item. It is measured and expressed through calculating the percentage of respondents answering each level for each item. This was calculated at baseline and follow-up. Change in response profiles between baseline and follow-up was analysed in groups which had improved or worsened in anchor measure scores. Analysis of the change in response profile provided an indication of which items were the "drivers" of change in the overall measure.

6.4.3.2. Non-weighted ICECAP score analysis.

The sum of the item scores for each participant was calculated without applying the preference weighting used to transform the item scores into a tariff score. This provides a non-value weighted ICECAP score. The correlation of this score with the anchor scores and change in this score by change groups in the anchor measures was analysed. Such an analysis of the non-weighted scores of a preference measure is an important step as it takes the question of whether a change is valued out of the analysis [259]. This non-weighted analysis allows the assessment of whether the descriptive system of the measure is responsive.

6.4.3.3. ICECAP value tariff analysis

The value weighted ICECAP tariff scores were calculated at baseline and follow-up and subsequently change scores were calculated. The correlation of the score with the anchor scores and the change in the score by change groups in the anchor measures was analysed. Comparison between the result of this analysis and the results of an analysis using the non-weighted scores, gives an indication of how adding the value of the scores to the measure might change the overall responsiveness.

6.4.4. Statistical analyses used in responsiveness analysis

6.4.4.1. Describing change

In this thesis, when describing the magnitude of change in ICECAP tariff score a standardised descriptive structure of "large", "moderate", "small" and "minor" changes was used to quantify change. These labels do not imply whether the change is important, meaningful or significant, rather they offer a structure to describe and discuss change. Minor changes are

defined as less than 0.01 changes, small changes are defined as change between 0.01 and 0.03, moderate changes are changes between 0.03 and 0.08 and large change were changes over 0.08.

Alongside actual change on a measure, change is reported as a percentage of possible change on the measures. This has the effect of standardising change and allowing comparison between non-weighted scores and tariff scores and between different measures. For example: the range of the tariff score of the ICECAP measure is 0 to 1, for the EQ-5D-3L index measure it is -0.59 to 1. The standardisation of change through the following equation allows the size of change to be compared between the two:

Change as a percentage of possible change = (change/range measure) * 100

6.4.4.2. Correlations

As described above, correlations were used to inform the selection of anchors for the analysis. They also provide a useful indication of the association between the anchor and ICECAP measure. Baseline and follow-up cross-sectional correlations and change correlations were calculated for all anchors. Pearson's correlation coefficient was used for data that were normally distributed and did not show skew [300]. Spearman's rank correlation coefficient was used for data that deviate from normality [301].

6.4.4.3. Paired t-test

The paired t-test was used to assess change between baseline and follow-up in the groups defined by the anchor values. The paired t-test tests the null-hypothesis that there has been no change in the mean response between baseline and follow-up and outputs indicate whether any change is statistically significant [307,308]. A weakness of the t-test is that it is highly

dependent on the sample size included in the measure. Therefore the sample size, which has little relation to the responsiveness of the measure, has a sizeable effect on the outcome of the test. The Wilcoxon rank test is the non-parametric equivalent of the t-test and was used when there was skew in the data.

6.4.4.4. Effect sizes and standardised response means

Two effect size statistics are reported: a standard effect size and the standardised response mean (SRM). Effect size statistics quantify the magnitude of change based on variation in the scores of the measure; simply put they are the ratio of signal to noise that exists within a measure. The standard effect size is calculated by dividing the change between baseline and follow-up by the standard deviation of the baseline scores [260].

$$ES = M_{Follow\text{-}up} - M_{Baseline} / SD_{Baseline}$$

The SRM is calculated by dividing the change between baseline and follow-up with the standard deviation of this change [260].

$$SRM = M_{Follow\text{-}up} - M_{Baseline} / SD_{Change}$$

Cohen [309] recommended the cut off values of 0.2, 0.5 and 0.8 are used to define very small, small, medium and large effect sizes [260] [264]. While some minor issues have been raised over accuracy when applying these values to SRMs [310], these values will be used here when categorising effect size and SRM changes.

6.4.4.5. Adding context through use of a reference measure

The EQ-5D-3L was used at a number of points as a reference measure that enabled greater context to be placed around the change observed in ICECAP measures scores. When it was

used as a reference measure to the non-weighted ICECAP scores, a non-weighted EQ-5D-3L score is used; when it was used as a reference measure with the value weighted ICECAP tariff scores, the preference weighted EQ-5D-3L index score was used. The use of the EQ-5D-3L as a comparator was not designed to assess which measure performs "best", as they are measures of two different constructs. Rather it was designed to increase the understanding of the size of changes in ICECAP scores in the context of another value (or preference) based measure.

6.5. Chapter summary

This chapter has described the process and the outcome of the recruitment of trials which provided data for the quantitative analyses in this thesis. The methods through which construct validity and responsiveness to change were assessed are described in detail, along with the statistical tests used. The results of the construct validity assessment are reported in Chapter 7 and the responsiveness results are reported in Chapter 8. The strengths and weakness of the methods used are examined in Chapter 9.

CHAPTER 7. QUANTITATIVE STUDY OF THE VALIDITY AND RESPONSIVENESS OF THE ICECAP MEASURES: VALIDITY RESULTS

7.1. Introduction

This chapter reports the results of two separate investigations into the construct validity of the ICECAP-A and ICECAP-O capability measures using baseline data taken from the BEEP trial and the PastBP trial. The chapter first reports the results from validation alongside the BEEP trial and then continues to report the results from validation alongside the PastBP trial.

Described in Chapter 6 are a number of *a priori* hypotheses that were formed. When an analysis was testing an *a priori* hypothesis, the results of the analysis are presented with contextual reference to the hypothesis. This is done both through discussion in text and the use of bold highlighting in tables to indicate that *a priori* an association was expected. Where *a priori* hypotheses were not formed, this is noted and the analysis is then considered exploratory. The level of support provided for the hypotheses as a whole is summarised at the end of the section for each trial.

7.2. BEEP trial

The characteristics of the BEEP trial participants are presented in Table 9. Results show the cohort to be older on average than the general population, with a roughly equal percentage of male and female participants. The psychological health of the cohort, as indicated by the Generalised Anxiety Disorder Assessment (GAD-7) and Patient Health Questionnaire depression scale (PHQ-8), was below the level indicating psychological disorders [295,296]. The average capability of the cohort, as indicated by the ICECAP-A tariff score was higher than values previously found in the general population [116]. The range of scores for both psychological health and capability, indicate that while the average scores indicate high levels of both construct, some participants were in very poor states. The EQ-5D-3L index score was below the national norm for people of this age group, which indicates worse than average population health [311]. The WOMAC scales indicated some participants have notable physical impairment and ill health. An item-by-item breakdown of the EQ-5D-3L scores (Appendix 21) show that reduced EQ-5D-3L scores were primarily due to increased pain and reduced mobility. This could reasonably be expected for a cohort of patients suffering from knee pain.

Table 9: Characteristics of BEEP trial participants

Characteristic	Mean values (SD)	Measure range	Sample range	Sample size
Socio-demographic				
Age (SD)	63.3 (9.9)		45 to 90	456
Gender (%male)	50.2%			456
Health and functioning				
ICECAP-A tariff	0.88 (0.12)	0.0 to 1.0	0.34 to 1.0	452
EQ-5D-3L index	0.63 (0.24)	-0.59 to 1.0	-0.18 to 1.0	442
WOMAC pain	8.5 (3.5)	0 to 20	0 to 18	449
WOMAC stiffness	3.8 (1.7)	0 to 8	0 to 8	451
WOMAC functioning	28.7 (12.2)	0 to 68	0 to 62	446
GAD-7	3.4 (4.7)	0 to 21	0 to 21	439
PHQ-8	4.1 (4.8)	0 to 24	0 to 24	442

526 patients were randomised and completed the baseline questionnaire packs. As noted in the methods, due to an administrative error, some participants were given an early version of the ICECAP-A questionnaire in their baseline questionnaire packs. To ensure rigour those participants were removed from this analysis, resulting in the results from 456 participants being used in this analysis.

7.2.1. ICECAP-A missing values

The BEEP trial baseline questionnaire packs were completed by participants at home and returned via post. Included within the baseline questionnaire was the consent form for the trial. Only those participants who had completed the consent form in this questionnaire pack were included in the analysis of missing values.

The number of missing values is in part dependent on the setting in which the questionnaire is completed and the protocol for ensuring completion. The EQ-5D-3L missing values allows a useful comparison. The EQ-5D-3L is a questionnaire of comparable patient burden (both are short form questionnaires consisting of five items) and it immediately preceded the ICECAP-A in the baseline questionnaire packs for this study. ICECAP-A tariff scores could not be calculated for four of the 456 participants in the trial, meaning that 452 (99%) of informants completed all items on the ICECAP questionnaire. The index scores of the EQ-5D-3L could not be calculated for 14 participants, meaning that 442 (97%) of informants completed all items on the EQ-5D-3L.

An item-by-item analysis (Table 10) shows the distribution of missing values by ICECAP and EQ-5D-3L items. There is no indication from the results of any item on the measure producing an unexpectedly high number of missing values, or notably more missing values than other items.

Table 10: Missing values by ICECAP-A items (EQ-5D-3L used as comparator) in BEEP trial.

Item	Values missing (%)
ICECAP –A	
Stability	4 (0.88%)
Attachment	2 (0.44%)
Autonomy	2 (0.44%)
Achievement	2 (0.44%)
Enjoyment	2 (0.44%)
EQ-5D-3L	
Mobility	2 (0.44%)
Self-care	2 (0.44%)
Usual activities	6 (1.32%)
Pain/Discomfort	9 (1.98%)
Anxiety/Depression	4 (0.88%)

7.2.2. ICECAP-A response patterns

Figure 13 shows the distribution of ICECAP-A tariff scores. The distribution has a significant skew ($p<0.01$) and kurtosis ($p<0.01$), with the highest concentration of response in scoress over 0.9.

Figure 13: Frequency distribution of ICECAP-A tariff scores at baseline in BEEP trial

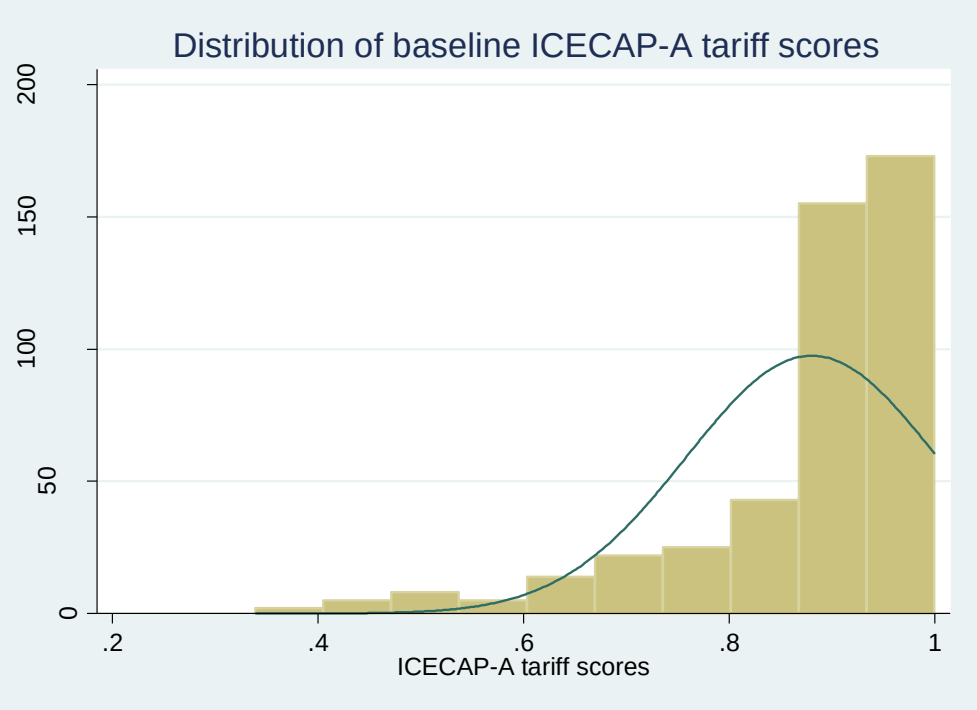


Table 11 and Figure 14 show the distribution of responses by item on the ICECAP-A. The frequency and percentage of participant answers for each level of each item are provided. Results show a very low number of respondents answering the bottom level of capability for any of the items, with the majority of respondents answering the top two capability levels. Over 60% of participants reported full capability (level 4) for Attachment and Autonomy, indicating the possibility that these items may suffer from ceiling effects in this population.

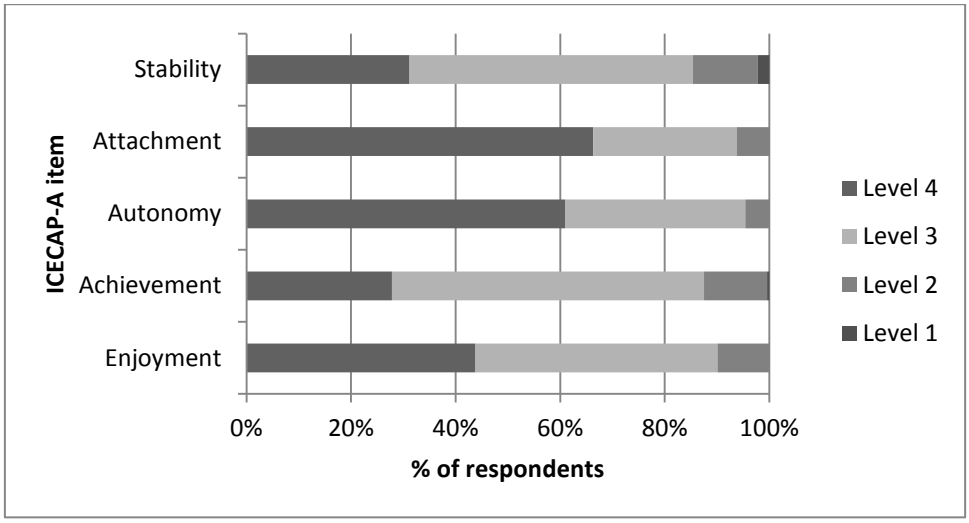
A breakdown of the ICECAP-A response profile from Al-Janabi et al's [116] investigation of construct validity in the general population is provided to allow for comparison. The pattern of response is similar to that seen in Al-Janabi's [116] results using the general population. The order of most selected to least selected responses for each item were the same for both samples (e.g. for the Stability item in both populations, the most frequently selected response level was the second, followed (in order) by the top, third and fourth levels), although in comparison to Al-Janabi et al's work [116] there was a notably larger percentage of BEEP trial participants who selected the top level of capability on the Autonomy and Achievement items.

Table 11: Distribution of responses by ICECAP-A item in BEEP trial (with Al-Janabi et al [116] values presented for comparison) (n=454)

Attribute and level	Frequency (%)	Al-Janabi (2012)
Stability*		
I am able to feel settled and secure in all areas of my life	141 (31.2%)	29%
I am able to feel settled and secure in some areas of my life	245 (54.2%)	51%
I am able to feel settled and secure in a few areas of my life	56 (12.4%)	17%
I am unable to feel settled and secure in any areas of my life	10 (2.2%)	3%
Attachment		
I can have a lot of love, friendship and support	301 (66.3%)	60%
I can have quite a lot of love friendship and support	125 (27.5%)	31%
I can have a little love, friendship and support	28 (6.2%)	7%
I cannot have any love friendship and support	0 (0.0%)	1%
Autonomy		
I am able to be completely independent	277 (61.0%)	47%
I am able to independent in many things	156 (34.4%)	41%
I am able to be independent in a few things	21 (4.6%)	11%
I am unable to be independent	0 (0.0%)	1%
Achievement		
I can achieve and progress in all aspects of my life	126 (27.8%)	18%
I can achieve and progress in many aspects of my life	271 (59.7%)	53%
I can achieve and progress in a few aspects o my life	55 (12.1%)	26%
I cannot achieve and progress in any aspects of my life	2 (0.4%)	2%
Enjoyment		
I can have a lot of enjoyment and pleasure	199 (43.8%)	37%
I can have quite a lot of enjoyment and pleasure	210 (46.3%)	46%
I can have a little enjoyment and pleasure	45 (9.9%)	15%
I cannot have any enjoyment and pleasure	0 (0.0%)	2%

* Stability sums to less than 454 due to missing values

Figure 14: Response profile of the ICECAP-A in BEEP trial



7.2.3. Socio-demographic variables and the ICECAP-A

The age and gender of the participants were available for all 456 participants used in this analysis. This section presents the analysis of the ICECAP-A's association with socio-demographic variables

7.2.3.1. Gender

There was no statistically significant association between the ICECAP-A tariff score and the gender of the participant. An item-by-item analysis (Table 12) showed that there were no statistically significant associations between any items of the ICECAP-A measure and the gender of the informant. These results were in agreement with hypotheses of no association between participant gender and ICECAP-A tariff score or an of the ICECAP-A items.

Table 12: Associations between gender and ICECAP-A items (n=452)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
Gender	0.722	0.799	0.417	0.439	0.843

* Significant at the 5% level, **Significant at the 1% level.

7.2.3.2. Age

Table 14 shows a negligible, non-significant correlation between the age of the participant and the ICECAP-A tariff score. Table 13 shows significant associations between age and Autonomy at the 1% level of significance and between age and Achievement at the 5% significance level. These results provide mixed support for *a priori* hypotheses. As hypothesised, age was associated with the Autonomy item. The expected association between age and ICECAP-A tariff score was not found. An unexpected association was seen between age and Achievement.

Table 13: Associations between age and ICECAP-A items (n=452)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
Age	0.103	0.481	**0.005****	0.036*	0.581

* Significant at the 5% level, **Significant at the 1% level.

Table 14: Correlations between measures in BEEP trial at baseline

	ICECAP-A tariff score	Age	EQ-5D-3L index score	WOMAC pain score	WOMAC stiffness score	WOMAC functioning score	IPQ score	GAD-7 score	PHQ -8 score	No. of co-morbidities
ICECAP-A tariff score	1.00									
Age	0.047	1.00								
EQ-5D-3L index score	0.486**	0.094*	1.00							
Womac pain	-0.146**	0.089	-0.455**	1.00						
Womac stiffness	-0.158**	0.136**	-0.45**	0.594**	1.00					
Womac functioning	-0.274**	0.166**	-0.596**	0.795**	0.674**	1.00				
IPQ score	-0.272**	-0.06	-0.447**	0.467**	0.37**	0.519**	1.00			
GAD-7 score	-0.517**	-0.047	-0.367**	0.184**	0.146**	0.192**	0.227**	1.00		
PHQ-8 score	-0.517*	-0.083	-0.427**	0.224**	0.233**	0.27**	0.271**	0.703**	1.00	
No. of co-morbidities	-0.22**	0.124**	-0.276**	0.141**	0.175**	0.205**	0.114*	0.204**	0.293**	1.00

* Significant at the 5% level, **Significant at the 1% level.

Table 14 shows correlations between all measures administered in the baseline questionnaire pack. The majority of correlations are weak to moderate in strength and all of the moderate correlations, and some of the weak correlations, are significant at the 1% level. Stronger correlations are seen within the WOMAC sub-scores of pain, stiffness and functioning. The correlations of the WOMAC sub-scales with the EQ-5D are stronger, than the correlations with the ICECAP. The measures of anxiety (GAD-7) and depression (PHQ-8) show stronger correlations with the ICECAP-A tariff than with measures of physical health.

7.2.4. Physical health and the ICECAP-A

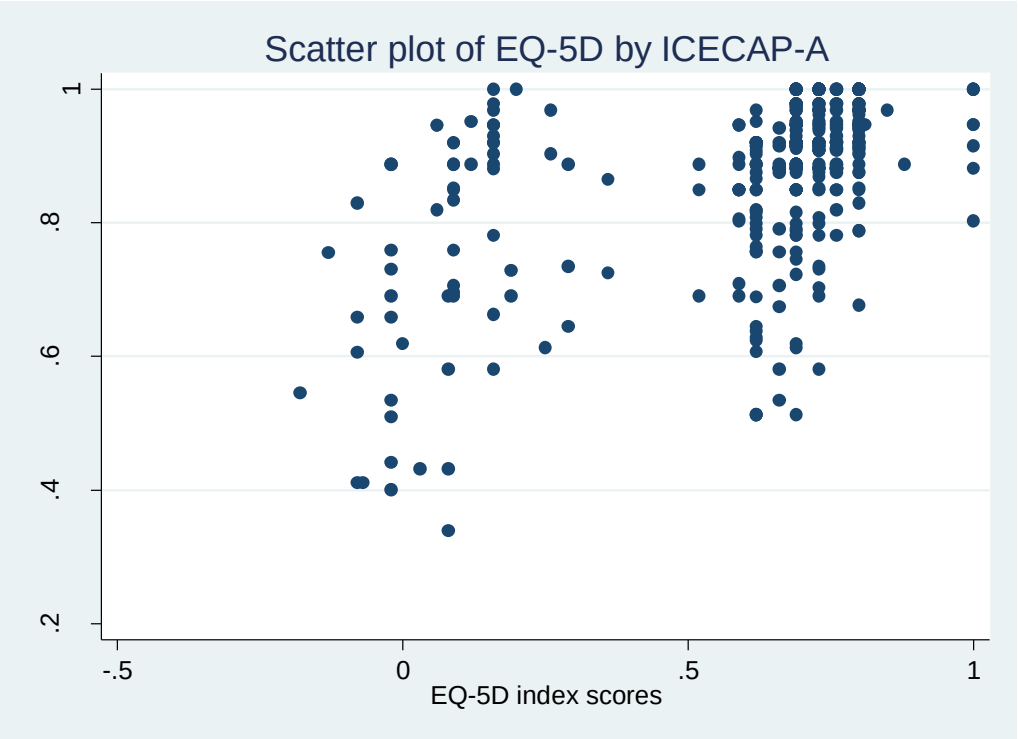
Four measures assessing physical health were completed by BEEP participants at baseline: the EQ-5D-3L, the WOMAC, the IPQ and participant co-morbidities. This section presents the analysis of the ICECAP-A's association with the EQ-5D-3L, the WOMAC and the patient co-morbidities. The analysis using the IPQ is placed in Appendix 22 and summarised briefly here. Where appropriate, comparisons are drawn with the *a priori* hypotheses reported in the methods chapter.

7.2.4.1. The EQ-5D-3L

7.2.4.1.1. ICECAP-A tariff score and the EQ-5D-3L

Table 14 shows there was a moderate, statistically significant correlation of 0.486 between the EQ-5D-3L index score and the ICECAP-A tariff scores. Figure 15 provides a graphical representation of this association. This suggests that as EQ-5D-3L scores increase, so do the ICECAP-A scores. The statistical significance of the correlation suggests this is unlikely to be due to chance.

Figure 15: Scatter plot of association between EQ-5D-3L index and ICECAP-A tariff



This moderate correlation, which was hypothesised *a priori*, between the ICECAP-A tariff and the EQ-5D-3L index score is suggestive of a non-perfect association between the two measures. Cross-sectional correlations, such as this, are not able to confirm causality. However, this moderate correlation indicates that ICECAP-A capability scores are dependent (at least in part) on determinants not measured by the EQ-5D-3L descriptive system. To assess whether this correlation held across different levels of health, correlations were assessed for those in the top quartile of EQ-5D-3L scores, the middle two quartiles of EQ-5D-3L scores combined, and the bottom quartile of health. Table 15 shows that the correlation was moderate at the lower levels of health and weak at the higher levels of the EQ-5D-3L health state.

Table 15: Correlations between ICECAP-A tariff and EQ-5D-3L index at differing levels of health (n=442)

	EQ-5D-3L ≥0.76	EQ-5D-3L <0.76 & ≥0.62	EQ-5D-3L <0.62
ICECAP-A tariff	0.238*	0.392**	0.407**

* Significant at the 5% level, **Significant at the 1% level.

Table 16 and Table 17 should be considered together. Table 16 shows that the ICECAP-A tariff score was significantly associated with each EQ-5D-3L item. Table 17 displays mean ICECAP-A tariff scores by EQ-5D-3L item levels. The results indicate a positive association between the two measures, across all EQ-5D-3L items. Higher scores on each EQ-5D-3L item was associated with higher ICECAP-A tariff scores. This finding is in line with the positive correlation between ICECAP-A tariff scores and EQ-5D-3L index presented in Table 14 and is supportive of the hypothesis of positive associations between ICECAP-A tariff and all EQ-5D-3L items.

Table 16: Associations between ICECAP-A tariff and EQ-5D-3L items (n=442)

Comparator	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
ICECAP-A tariff	**<0.001****	**<0.001****	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

In Table 17, with the one exception of pain and discomfort, differences in the mean ICECAP-A tariff score between the top two levels were statistically significant. For three items - usual activities, pain/discomfort and anxiety and depression - there were significant differences in the ICECAP-A scores between the bottom two levels. The exceptions were mobility, where no participant was bedbound, and self-care, where the power of the test suffered from low numbers in the bottom group. These results provide strong evidence that, in a medium sized

sample, the ICECAP-A tariff score differs across the levels of the health dimensions included in the EQ-5D-3L.

Table 17: Mean ICECAP-A tariff score by EQ-5D-3L item levels (n=451)

EQ-5D-3L Attribute	ICECAP-A tariff score (95% CI)
Mobility	
No problems (n=145)	0.91 (0.903, 0.930)
Some problems (n=306)	0.86 (0.848, 0.878)†
Bedbound (n=0)	-
Self-care	
No problems (n=411)	0.89 (0.884, 0.904)
Some problems (n=39)	0.74 (0.681, 0.794)†
Unable (n=1)	0.76 (n/a)
Usual activities*	
No problems (n=194)	0.92 (0.909, 0934)
Some problems (n=246)	0.85 (0.839, 0.871)†
Unable (n=7)	0.67 (0.491, 0.844)††
Pain and discomfort*	
No pain and discomfort (n=11)	0.94 (0.895, 0.977)
Moderate pain and discomfort (n=380)	0.89 (0.882, 0.903)
Extreme pain and discomfort (n=54)	0.78 (0.733, 0.829)††
Anxiety and depression*	
Not anxious or depressed (n=321)	0.92 (0.912, 0.929)
Moderately anxious or depressed (n=115)	0.8 (0.775, 0.826)†
Extremely anxious or depressed (n=13)	0.58 (0.501, 0.662)††

Statistically significant differences at the 1% level in the mean ICECAP-A tariff score between †some(moderate)/no(not) response options and ††unable(extreme)/some(moderate) levels of the EQ-5D-3L. *These items sum to less than 451 due to missing values.

7.2.4.1.2. ICECAP-A items and the EQ-5D-3L

Table 18 shows associations between ICECAP-A items and the EQ-5D-3L index scores. Results show that all items were significantly associated with the EQ-5D-3L index score. As can be seen, an unexpected association between the EQ-5D-3L index and the ICECAP-A item of Attachment was found.

Table 18: Associations between EQ-5D-3L index and ICECAP-A items (n=422)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
EQ-5D-3L index score (n=442)	**<0.001****	<0.001**	**<0.001****	**<0.001****	**<0.001****

Table 19 should be considered together with Table 18. The results show that EQ-5D-3L index scores were strongly and positively related to all ICECAP-A attributes. There were statistically significant differences in the EQ-5D-3L index score between the top and the second levels and the second and third levels on every ICECAP-A attribute. No statistically significant differences were seen between the scores of the bottom two levels of any attribute, which may be due to small numbers resulting in poor statistical power of the Kruskal-Wallis multiple group comparisons at these lower levels.

Table 19: Mean EQ-5D-3L index scores by ICECAP-A item levels (n=442)

ICECAP-A Attribute	Mean EQ-5D-3L score (95% CI)
Stability*	
I am able to feel settled and secure in all areas of my life (n=137)	0.69 (0.656, 0.718)††
I am able to feel settled and secure in many areas of my life (n=238)	0.66 (0.637, 0.684)†
I am able to feel settled and secure in a few areas of my life (n=55)	0.42 (0.336, 0.512)
I am unable to feel settled and secure in any areas of my life (n=10)	0.12 (-0.073, 0.317)
Attachment	
I can have a lot of love, friendship and support (n=292)	0.65 (0.625, 0.676)††
I can have quite a lot of love, friendship and support (n=123)	0.60 (0.563, 0.646)†
I can have a little love, friendship and support (n=27)	0.45 (0.321, 0.579)
I cannot have any love, friendship and support (n=0)	-
Autonomy	
I am able to be completely independent (n=270)	0.69 (0.671, 0.713)††
I am able to be independent in many things (n=152)	0.56 (0.516, 0.599)†
I am able to be independent in a few things (n=20)	0.24 (0.092, 0.389)
I am unable to be at all independent (n=0)	-
Achievement	
I can achieve and progress in all aspects of my life (n=122)	0.71 (0.682, 0.738)††
I can achieve and progress in many aspects of my life (n=264)	0.64 (0.613, 0.664)†
I can achieve and progress in a few aspects of my life (n=54)	0.38 (0.294, 0.473)
I cannot achieve and progress in any aspects of my life (n=2)	0.3 (n/a)
Enjoyment	
I can have a lot of enjoyment and pleasure (n=193)	0.70 (0.678, 0.726)††
I can have quite a lot enjoyment and pleasure (n=205)	0.60 (0.570, 0.636)†
I can have a little enjoyment and pleasure (n=44)	0.39 (0.294, 0.491)
I cannot have any enjoyment and pleasure (n=0)	-

Statistically significant differences at the 1% level in the mean EQ-5D-3L index score between †† all (a lot)/many (quite a lot) and † many (quite a lot)/few (a little) ICECAP-A response options. *Stability sums to less than 442 due to missing values.

Table 20 shows the associations between the ICECAP-A items and the EQ-5D-3L items. The associations which were expected *a priori* are highlighted in bold. As can be seen, all of the 19 hypothesised associations were significant at the 5% statistical level, with all but one significant at the 1% level. Four associations were found that were unexpected: Attachment

was unexpectedly associated with self-care, usual activities and anxiety, while Stability was unexpectedly associated with usual activities.

Table 20: Associations between ICECAP-A and EQ-5D-3L items (n=442)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
EQ-5D-3L					
Mobility	**0.038***	0.089	**<0.001****	**<0.001****	**<0.001****
Self-care	**<0.001****	0.002*	**<0.001****	**<0.001****	**<0.001****
Usual activities	<0.001**	0.022*	**<0.001****	**<0.001****	**<0.001****
Pain	**<0.001****	0.264	**<0.001****	**<0.001****	**<0.001****
Anxiety	**<0.001****	<0.001**	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level. Hypothesised associations are highlighted in bold font

Evidence of the direction of the relationship between the ICECAP-A items and the EQ-5D-3L items is provided in Table 21 using Goodman and Kruskal's Gamma. All rank correlations were in the anticipated direction. All correlations were negative due to the difference in scoring between the ICECAP-A and the EQ-5D-3L (i.e. the top level on ICECAP items is 4, whereas the top level for EQ-5D-3L items is 1). Therefore, the results indicate that high scores on the health attribute (e.g. mobility) associates with high scores on the capability attribute (e.g. Autonomy).

The associations that were hypothesised *a priori* to be particularly strong are highlighted in bold. All these associations were moderate or strong. There were unexpectedly strong correlations seen between 1) Pain/Discomfort and Autonomy and 2) Anxiety and Depression and Stability and Achievement.

Table 21: Rank correlations between ICECAP-A and EQ-5D-3L items (n=442)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
EQ-5D-3L					
Mobility	-0.178	-0.157	**-0.485**	-0.396	-0.342
Self-care	-0.563	-0.457	**-0.868**	-0.779	-0.514
Usual Activities	-0.368	-0.250	**-0.641**	-0.623	-0.460
Pain/Discomfort	-0.400	-0.226	-0.545	-0.475	**-0.497**
Anxiety/Depression	-0.836	-0.480	-0.613	-0.773	**-0.724**

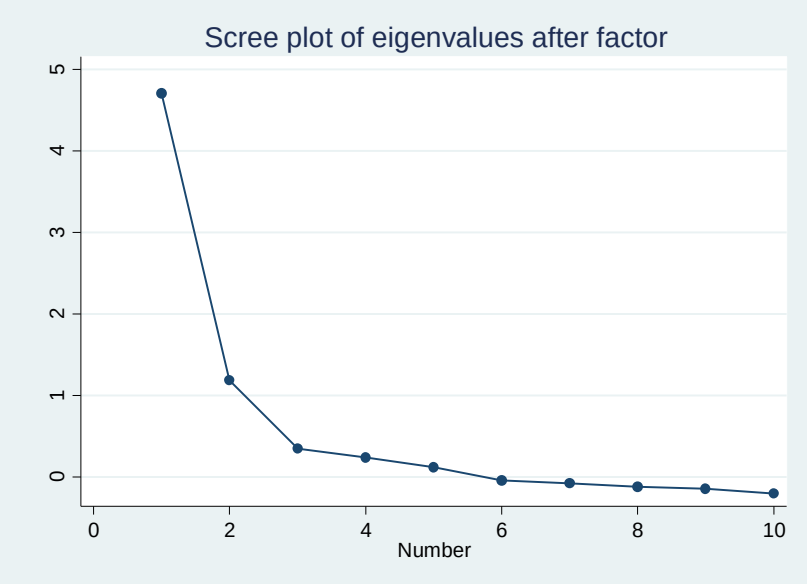
Highlighted in bold are the associations which were hypothesised to the strongest for each EQ-5D-3L item.

7.2.4.1.3. Exploratory factor analysis

The results presented above indicate a closer association between the ICECAP-A measure and the EQ-5D-3L than was hypothesised. There were a number of unexpected associations between items (Table 20) and the strength of some of these associations was stronger than anticipated (Table 21). This raises an important question: are the ICECAP-A and EQ-5D-3L measuring different constructs, or are they different measures of the same construct? To assess this, an exploratory factor analysis was completed using the 5 items from both measures.

The number of factors retained in the final solution was chosen with reference to the Kaiser Criterion: both the scree plot (see Figure 16) and the Eigen Values indicated that maintaining 2 factors was the optimal solution.

Figure 16: Factor analysis eigen values by factor number



The choice of an oblique promax rotation was confirmed as correct through use of the STATA "estat common" command (post rotation) which indicated a correlation of -0.5149 between the factors. This correlation is stronger than 0.32 which is considered the point at which oblique rotations are appropriate.

Table 22: Exploratory factor analysis comparing the ICECAP-A and EQ-5D-3L items (n=442)

	Rotated item loading on a 2-factor solution.	
	Factor 1	Factor 2
EQ-5D-3L		
Mobility		0.816
Self-Care		0.790
Usual Activities		0.693
Pain		0.667
Anxiety and Depression	-0.741	
ICECAP-A		
Stability	0.859	
Attachment	0.671	
Autonomy	0.398	-0.459
Achievement	0.677	-0.224
Enjoyment	0.825	
Factor correlations	-0.515	

Loadings of <0.2 are dropped to allow easy interpretation of results.

Table 22 shows a two factor solution indicating that two separate, but correlated attributes are assessed by the pooled items of EQ-5D-3L and the ICECAP-A. The majority of EQ-5D-3L items (Mobility, Self-care, Usual Activities and Pain) loaded strongly onto factor two, while the majority of ICECAP-A items (Stability, Attachment, Achievement, Enjoyment) and one EQ-5D-3L item (Anxiety and Depression) loaded strongly onto factor one. The loading of Autonomy split equally, with moderate loadings onto both factors. Therefore these results strongly indicate that the ICECAP-A and EQ-5D-3L are largely capturing different constructs.

7.2.4.2. WOMAC

Table 14 shows that, in line with hypothesised associations, all three of the WOMAC scales showed weak negative, but statistically significant correlations with the ICECAP-A tariff

score. When levels of pain, stiffness and functioning problems due to knee pain are high in an individual; it seems to be reflected to some degree in lower ICECAP-A tariff scores.

The results in Table 23 indicate that each sub-scale of the WOMAC showed a statistically significant association with the four ICECAP-A items of Stability, Autonomy, Achievement and Enjoyment at the 1% level of statistical significance. No statistically significant associations were seen between the Attachment item and any of the WOMAC sub-scales. *A priori* expected associations are highlighted in bold and as can be seen it was expected that all ICECAP-A items except Attachment would be associated with each of the WOMAC sub-scales. All expected associations were significant at the 1% level.

Table 23: Associations between WOMAC subscales and ICECAP-A items

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
WOMAC pain (n=441)	**0.001****	0.353	**<0.001****	**<0.001****	**<0.001****
WOMAC stiff (n=447)	**0.01***	0.789	**<0.001****	**0.002****	**0.008****
WOMAC func (n=389)	**<0.001****	0.07	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

7.2.4.3. The Brief Illness Perception Questionnaire summary

The results of the IPQ analysis are presented in Appendix 22 and a summary is provided here. The IPQ showed a weak statistically significant association with the ICECAP-A tariff. Four items of the questionnaire showed a statistically significant association with the ICECAP-A tariff. These were: affect - an assessment of the affect that knee pain has on life; symptoms - an assessment of how many symptoms are experienced; concern - an assessment of the level of concern over knee pain; and emotion - an assessment of the emotional effect of knee pain.

There were a number of statistically significant associations between IPQ items and ICECAP-A items.

7.2.4.4. Co-morbidities

A ten item questionnaire of co-morbidities was included in the baseline questionnaire pack of the BEEP trial. This questionnaire contained the question "Have you ever been told that you have any of the following?" and the options were: high blood pressure, angina, heart failure, stroke, depression, osteoporosis, diabetes, asthma, bronchitis and heart attack. The first part of this analysis presents a summed value of "total number co-morbidities" and the second part of the analysis looks at the impact of individual co-morbidities on the ICECAP-A tariff scores. This analysis was exploratory.

Table 14 shows a weak, statistically significant correlation of -0.22 between the ICECAP-A scores and the number of co-morbidities reported by participants. This correlation indicates that the more co-morbidities a person has, the lower their ICECAP-A scores tend to be. To analyse this further the total number of co-morbidities was transformed into a categorical variable of none, some (1 to 2) and many (2 to 10). Table 24 shows the mean ICECAP scores for participants in each of these groups. The groups with less co-morbidities had higher ICECAP-A scores.

Table 24: Mean ICECAP-A tariff scores by prevalence of co-morbidities (n=229)

	ICECAP-a tariff score (95% CI)
None	0.903 (0.884, 0.921)
Some (1 to 2 co-morbidities)	0.883 (0.869, 0.897)†
Many (3 or more co-morbidities)	0.815 (0.772, 0.858)

† Statistically significant difference between Some and Many co-morbidities

Table 25 shows the associations between the categorical co-morbidities variable and each ICECAP-A item. Co-morbidities are associated with each ICECAP-A item apart from Attachment. As stated above, this analysis was exploratory, therefore no *a priori* hypotheses were formed. However, for other health measures, such as the EQ-5D-3L and WOMAC, indicators of health were hypothesised to associate with all ICECAP-A items apart from Attachment. Such a pattern can be seen with the comorbidities question.

Table 25: Associations between co-morbidities categorical variable and ICECAP-A items (n=229)

	Stability	Attachment	Autonomy	Achievement	Enjoyment
Comorbidities	0.001**	0.062	0.001**	0.001**	0.001**

* Significant at the 5% level, **Significant at the 1% level.

Table 26 shows each of the individual co-morbidities assessed in the BEEP trial and the mean ICECAP-A tariff score for those who did and did not report that comorbidity. Significant differences were found between the ICECAP-A tariff scores for those with and without depression and osteoporosis. A number of the analyses suffered from low numbers in the group reporting the comorbidity. This limitation applies particularly to angina, heart failure, stroke and heart attack.

Table 26: Mean ICECAP-A tariff scores by number of self-reported co-morbidities

	ICECAP-A mean tariff scores	
	Reported comorbidity (95% CI)	**Comorbidity not reported (95% CI)**
High blood pressure	0.878 (0.862, 0.894) (n=211)	0.883 (0.867, 0.899) (n=241)
Angina	0.847 (0.783, 0.912) (n=19)	0.882 (0.87, 0.893) (n=433)
Heart failure	0.869 (0.741, 0.998) (n=8)	0.881 (0.87, 0.892) (n=444)
Stroke	0.859 (0.815, 0.903) (n=15)	0.881 (0.87, 0.893) (n=437)
Depression**	0.805 (0.773, 0.837) (n=99)	0.902 (0.891, 0.912) (n=353)
Osteoporosis*	0.823 (0.772, 0.875) (n=30)	0.885 (0.873, 0.896) (n=422)
Diabetes	0.863 (0.829, 0.896) (n=62)	0.883 (0.871, 0.895) (n=390)
Asthma	0.851 (0.881, 0.891) (n=59)	0.885 (0.873, 0.896) (n=393)
Bronchitis	0.842 (0.786, 0.898) (n=32)	0.883 (0.872, 0.895) (n=420)
Heart attack	0.902 (0.855, 0.958) (n=16)	0.88 (0.868, 0.891) (n=436)

* Significant at the 5% level, **Significant at the 1% level.

7.2.5. Psychological health and the ICECAP-A

Two measures of psychological health were included in the BEEP trial: the GAD-7 and the PHQ-8. This section presents the ICECAP-A's association with these measures and comments on how the results compare to the *a priori* hypotheses.

7.2.5.1. GAD-7 and PHQ-8

Table 14 shows that the hypothesised negative associations with the ICECAP-A tariff score were seen for both the GAD-7 and the PHQ-8. The moderate associations of -0.517 for both measures is the strongest correlation seen between the ICECAP-A tariff score and the scores of any measure included in the baseline questionnaire pack of the BEEP trial.

Table 27 and Table 28 show that the GAD-7 and the PHQ-8 scores are associated with every ICECAP-A item. For both measures associations were found with Attachment and Achievement that were not hypothesised. This suggests that there was a closer association between the dimensions of the ICECAP-A measure and measures of anxiety and depression than was expected.

Table 27: Associations between GAD-7 score and ICECAP-A items (n=439)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
GAD-7	**<0.001****	<0.001**	**<0.001****	<0.001**	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

Table 28: Associations between PHQ-8 score and ICECAP-A items (n=442)

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
PHQ-8	**<0.001****	<0.001**	**<0.001****	<0.001**	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

The GAD-7 and PHQ-8 scores were split into categorical variables based on clinically meaningful cut-offs. For the GAD-7 a score equal to or greater than 8 indicates anxiety

disorder [295,297,312], while for the PHQ-8 a score equal to or greater than 10 indicates probable depression [296]. Table 29 shows the ICECAP-A tariff scores for those with and without anxiety and depression, as indicated by the GAD-7 and the PHQ-8. A difference of 0.16 was seen between those with and without probable anxiety disorder and a difference of 0.18 was seen between those with and without probable depression. The differences in ICECAP-A tariff scores between those with and without probable depression or anxiety disorder were significantly different. No *a priori* hypotheses were formed as to the association of the ICECAP-A with the categorical transformations of the GAD-7 and PHQ-8 measures. However, these results are largely in line the hypothesis that the ICECAP-A tariff score would be higher in those in good psychological health as judged by the GAD-7 and PHQ-8.

Table 29: Mean ICECAP-A tariff scores by GAD-7 and PHQ-8 levels (based on clinical cut-offs)

Measure of psychological health	**ICECAP-A tariff score (SD)**
GAD-7 (n=439)	
No anxiety disorder	0.91 (0.089)**
Probable anxiety disorder	0.75 (0.17)
PHQ-8 (n=442)	
Not depressed	0.91 (0.091)**
Probable depression	0.73 (0.16)

** Statistically significant differences at the 1% level in the mean ICECAP-A tariff score between No(not)/probable response options

7.2.6. Comparison of results with hypotheses

A priori hypotheses were formed and have been referred to at a number of points throughout the analysis. Table 30 presents a comparison of hypotheses and results. Taken as a whole these results provide positive, supportive evidence for the validity of the ICECAP-A measures. The majority of hypothesised associations were found; meaning that the measure behaved largely as expected. As can be seen from Table 30, confirmatory evidence pertaining to the hypotheses referring to the association of the ICECAP-A tariff with socio-demographic, physical health and psychological health variables was found.

No numerical strength of association or correlation was hypothesised as it was not felt to be possible to do this accurately. However, discussed in Chapter 3 was the expectation that strong correlations between the ICECAP-A and measures of physical and psychological health would not be found, and that moderate correlations would be expected as there are a number of influences upon capability other than health. As can be seen from the results above and Table 30, all correlations are moderate or less; no strong correlations were found.

Table 30: A comparison of hypotheses and results from the BEEP trial

Hypotheses				Results			
	Association	Direction	Strength		Association	Direction	Strength
Socio-demographic				**Socio-demographic**			
Age	Yes			Age	No		
Gender	No			Gender	No		
Physical health				**Physical health**			
EQ-5D-3L	Yes	Positive		EQ-5D-3L	Yes	Positive	Moderate
WOMAC Pain	Yes	Negative		WOMAC Pain	Yes	Negative	Weak
WOMAC stiffness	Yes	Negative		WOMAC stiffness	Yes	Negative	Weak
WOMAC functioning	Yes	Negative		WOMAC functioning	Yes	Negative	Weak
IPQ	*Exploratory analysis*			IPQ	Yes	Negative	Weak
Co-morbidities	*Exploratory analsyis*			Co-morbidities	Yes	Negative	Weak
Psychological health				**Psychological health**			
GAD-7	Yes	Negative		GAD-7	Yes	Negative	Moderate
PHQ-8	Yes	Negative		PHQ-8	Yes	Negative	Moderate

No association between gender and the ICECAP-A tariff score was expected; no significant association was seen. The item by item analysis showed mixed support for the hypotheses referring to age. The expected association of the tariff with age was not found. As expected, age showed a significant association with the ICECAP item measuring independence, the Autonomy item, and an unexpected association with the Achievement item.

The associations between measures of physical health in the BEEP trial and the ICECAP-A tariff score were as expected. It was hypothesised that a positive correlation would be found between the EQ-5D-3L index score and the ICECAP-A tariff score: a moderate positive correlation was found. It was expected that the scores of the WOMAC subscales of pain, stiffness and functioning would negatively associate with the ICECAP-A tariff (e.g.greater

pain, stiffness and functioning problems would be associated with lower capability). A weak negative association between each of the subscales and the tariff score was seen.

The item-by-item analysis of the ICECAP-A and the measures of physical functioning showed some additional non-hypothesised associations. As expected, statistically significant associations were seen between all three WOMAC sub-scales and all ICECAP-A items apart from Attachment. All hypothesised associations between ICECAP-A and EQ-5D-3L items were found and an additional four unexpected associations were found. This discordance between the hypotheses and the results indicate a closer than expected association between the ICECAP-A items and the EQ-5D-3L items.

The associations and correlations between measures of psychological health and the ICECAP-A tariff score were as expected. A negative correlation was expected *a priori*; a moderate negative correlation was found. The correlations with the GAD-7 and the PHQ-8 were the strongest association of any variable with the ICECAP-A tariff score. The item-by-item analysis showed some additional associations which were not expected a priori, suggesting a closer association than expected.

7.3. PastBP trial

The characteristics of the PastBP trial participants are presented in Table 31. Results show this to be an elderly, majority male sample. The average capability of participants, as indicated by an average ICECAP-O score of 0.85, was comparable to values found in the general population [118]. The range of ICECAP-O scores indicated that some participants had low capability scores. Participants were on average in reasonable physical health. The mean EQ-5D-3L score for this population was comparable to the average score for this age group [311]. The average number of symptoms and side-effects experienced by participants was 6 out of the 24 assessed. The distribution and sample range of both the EQ-5D-3L scores and the symptom and side-effects score show a small number of participants to be in very poor health states. The cognitive functioning of the sample was above the MMSE cut-off of 25, indicating good cognitive impairment [313]. Scores on the sub-scales of the SF-36 indicate that physical functioning, limitations to a person's role due to physical factors and vitality were the main impairments to health in this sample.

Table 31: Characteristics of PastBP trial participants

Attribute	Mean values (SD)	Measure range	Sample range	Sample size
Socio-demographic				
Age mean (SD)	71.2 (9.16)		41 to 91	529
Gender (% male)	59.4%			529
Health and functioning				
ICECAP-O score	0.85 (0.13)	0 to 1	0.25 to 1.0	456
EQ-5D-3L	0.72 (0.24)	-0.59 to 1	-0.07 to 1.0	476
EQ-5D-3L VAS	73.3 (18.0)	0 to 100	0 to 100	470
MMSE	27.9 (2.5)	0 to 30	0 to 30	472
Number of SSE	6	0 to 24	0 to 20	299
SF-36 sub-scale scores				
• Physical function	41.0 (12.3)	14.9 to 57.0	14.9 to 57.0	418
• Physical role	40.1 (17.4)	17.7 to 56.8	17.7 to 56.8	470
• Emotional role	43.8 (18.1)	9.2 to 55.9	9.2 to 55.9	462
• Vitality	48.7 (10.7)	20.9 to 70.8	20.9 to 70.8	457
• Mental health	50.5 (9.9)	14.5 to 64.1	14.5 to 64.1	472
• Social function	47.3 (11.2)	13.2 to 56.8	13.2 to 56.8	478
• Pain	47.9 (11.4)	19.8 to 62.1	19.8 to 62.1	474
• General health	44.9 (9.8)	16.2 to 63.9	16.2 to 63.9	454

7.3.1. ICECAP-O missing values

The majority of baseline questionnaire packs for the PastBP trial were completed in clinic under the supervision of a researcher. The primary role of the researcher was to confirm the diagnosis of the patient and facilitate blood pressure measurement and randomisation. Forms were self-completed. As noted above, the completion rates for a measure are dependent on methods of the trial, therefore the EQ-5D-3L is used as a comparator with similar patient burden.

During data entry, all measures were marked electronically as either "completed" or "not-completed". Completed meant that participants had answered one or more items on the measure. Not-completed meant that the measure was included in the baseline questionnaire pack administered to participants, but the participant had not answered any items on the measure. Table 32 shows slightly reduced completion rates for ICECAP-O in comparison to EQ-5D-3L. The difference in the total number of questionnaires is due a small number of the early baseline questionnaire packs not including ICECAP-O.

Table 32: Completion rates of ICECAP-O (EQ-5D-3L used as comparator) in PastBP trial

Measure	Completed (%)	Not-completed (%)	Total
ICECAP-O	460 (92.6)	37 (7.4)	497
EQ-5D-3L	484 (93.4)	34 (6.6)	518

Missing items were assessed by excluding "not-completed" questionnaires from the analysis and assessing the missing items on those questionnaires where at least one item on the measure had been completed. Table 33 shows a very low number of both ICECAP-A and EQ-5D-3L items were not completed by the participants who had attempted to complete the

measure. There is no indication of increased non-response on any of the items for either measure.

Table 33: Missing values by ICECAP-O items (EQ-5D-3L used as comparator) in PastBP trial

Item	Items missing (%)
ICECAP –0	
Attachment	2 (0.4)
Security	1 (0.2)
Role	3 (0.7)
Enjoyment	1 (0.2)
Control	1 (0.2)
EQ-5D-3L	
Mobility	3 (0.6)
Self-care	4 (0.8)
Usual activities	3 (0.6)
Pain/Discomfort	3 (0.6)
Anxiety/Depression	3 (0.6)

Analysis completed using participants who had completed at least one item on measure.

7.3.2. ICECAP-O response patterns

Figure 17 shows the distribution of ICECAP-O tariff scores. The distribution has a notable skew ($p<0.01$) and kurtosis ($P<0.01$), with the highest concentration being in values over 0.8.

Figure 17: Frequency distribution of ICECAP-O tariff scores at baseline in PastBP trial

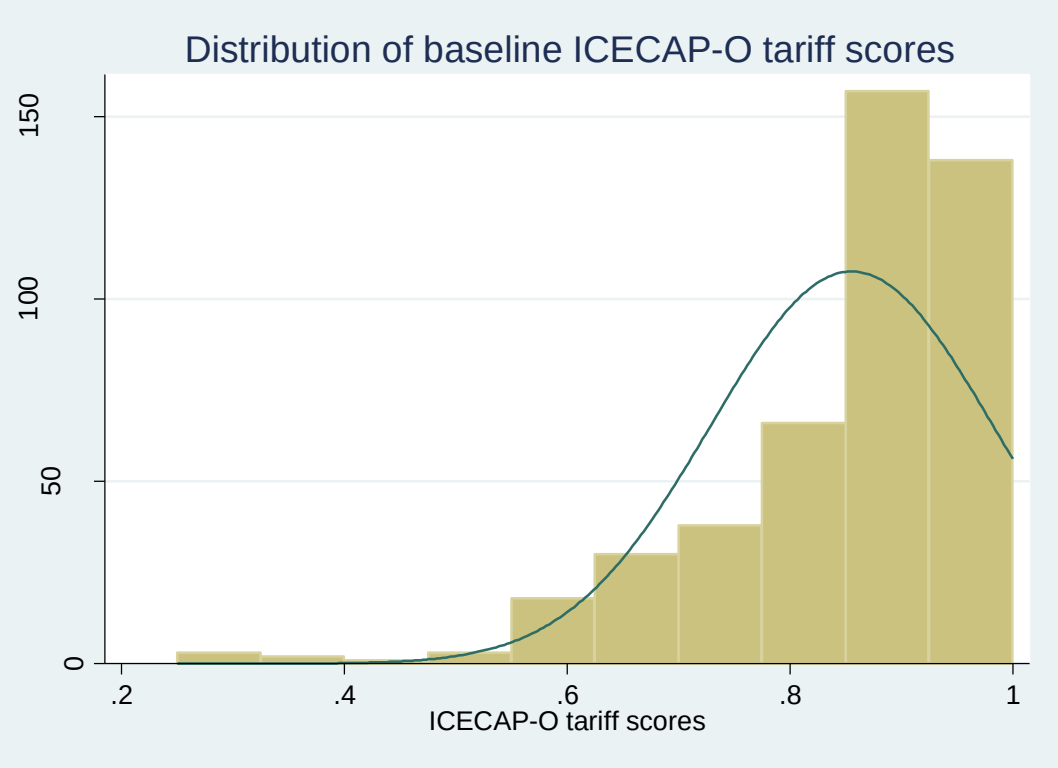


Table 34 and Figure 18 shows the frequency of responses for each level of each item of the ICECAP-O measure. The percentage of participants answering each level is given, alongside percentages from Coast et al's [117] investigation of the construct validity of the ICECAP-O measure in the general population to allow comparison.

Results show a low number of participants selecting the bottom level of capability. The item with the highest percentageselecting the bottom level was Security with 6.1%. The majority of respondents selected the top two levels of each item, with 60% in the top level for Attachment and 51% for Control. On all items apart from the Control item the order of the

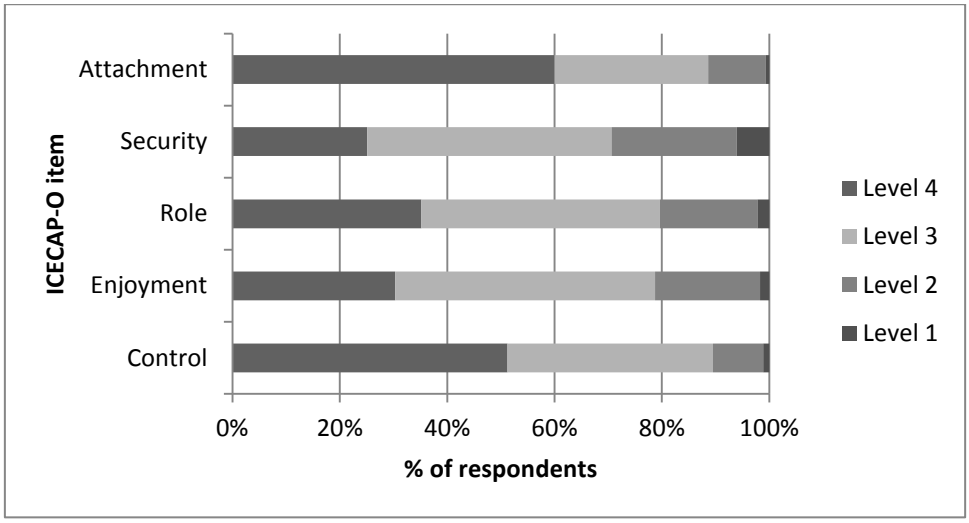
most selected to least selected levels is the same for both the PastBP and the Coast et al [117] general population sample. Notable increases in the percentage of participants who selected the top capability level for Role and Control was seen in comparison to the general population sample.

Table 34: Distribution of responses by ICECAP-O item in PastBP trial (Coast et al [117] presented for comparison) (n=459)

Attribute and Level	Frequency (%)	General pop percentage
Attachment*		
I can have all of the love and friendship that I want	275 (60.0%)	57%
I can have a lot of the love and friendship that I want	131 (28.6%)	30.2%
I can have a little of the love and friendship that I want	49 (10.7%)	8.9%
I cannot have any of the love and friendship that I want	3 (0.7%)	3.8%
Security		
I can think about the future without any concern	115 (25.1%)	18.7%
I can think about the future with only a little concern	209 (45.5%)	38.7%
I can only think about the future with some concern	107 (23.3%)	30.2%
I can only think about the future with a lot of concern	28 (6.1%)	12.4%
Role*		
I am able to do all of the things that make me feel valued	161 (35.2%)	26.1%
I am able to do many of the things that make me feel valued	203 (44.4%)	47.4%
I am able to do a few of the things that make me feel valued	83 (18.2%)	22.9%
I am unable to do any of the things that make me feel valued	10 (2.2%)	3.5%
Enjoyment		
I can have all of the enjoyment and pleasure that I want	139 (30.3%)	23.5%
I can have a lot of the enjoyment and pleasure that I want	222 (48.4%)	51.4%
I can have a little of the enjoyment and pleasure that I want	90 (19.6%)	21%
I cannot have any of the enjoyment and pleasure that I want	8 (1.7%)	4.1%
Control		
I am able to be completely independent	235 (51.2%)	38.7%
I am able to be independent in many things	176 (38.3%)	41.3%
I am able to be independent in a few things	43 (9.4%)	17.8%
I am unable to be at all independent	5 (1.1%)	2.2%

* Items sum to less than 459 due to missing values.

Figure 18: Response profile of ICECAP-O in PastBP trial



7.3.3. Socio-demographic variables and the ICECAP-O

A small amount of socio-demographic information was collected on the participants in the PastBP trial. The age and gender of participants was available for analysis. This section presents the analysis of the association of the ICECAP-O with these variables.

7.3.3.1. Gender

There was no statistically significant association between the ICECAP-O tariff score and the gender of the participant. An item-by-item analysis, presented in Table 35, shows there was no statistically significant association between any of the ICECAP-O items and gender. These results are in agreement with the *a priori* hypotheses which anticipated no associations.

Table 35: Associations between gender and ICECAP-O items (n=456)

Comparator	Attachment	Security	Role	Enjoyment	Control
Gender (459)	0.954	0.177	0.125	0.735	0.154

7.3.3.2. Age

Table 37 shows a negligible, non-significant correlation between age and the ICECAP-O tariff score. Table 36 shows that a statistically significant association at the 1% level exists between age and the ICECAP-O items of Security and Control. The association between Control and age was expected, while the association between Security and age was unexpected. These results therefore provide mixed support for the *a priori* hypotheses.

Table 36: Associations between age and ICECAP-O items (n=456)

Comparator	Attachment	Security	Role	Enjoyment	Control
Age	0.608	0.001**	0.969	0.440	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

Table 37: Correlations between measures in PastBP trial at baseline

	ICECAP-O tariff	Age	EQ-5D-3L index	EQ-5D-3L VAS	Number of SSE	Physical function	Physical role	Emotional role	Vitality	Mental Health	Social function	Pain	General health
ICECAP-O tariff	1.00												
Age	-0.013	1.00											
EQ-5D-3L index score	0.518**	-0.125**	1.00										
EQ-5D-3L VAS	0.533**	-0.073	0.536**	1.00									
Number of SSE	-0.477**	-0.035	-0.547**	-0.499**	1.00								
Physical function	0.467**	-0.269**	0.685**	0.541**	-0.507**	1.00							
Physical role	0.435**	-0.157**	0.557**	0.533**	-0.465**	0.592**	1.00						
Emotional role	0.436**	-0.051	0.399**	0.389**	-0.364**	0.394**	0.559**	1.00					
Vitality	0.599**	-0.041	0.604**	0.646**	-0.579**	0.572**	0.581**	0.482**	1.00				
Mental Health	0.521**	-0.140**	0.382**	0.461**	-0.438**	0.301**	0.330**	0.494**	0.585**	1.00			
Social function	0.527**	-0.024	0.589**	0.547**	-0.453**	0.558**	0.608**	0.561**	0.607**	0.498**	1.00		
Bodily pain	0.328**	0.007	0.696**	0.479**	-0.483**	0.508**	0.486**	0.358**	0.480**	0.310**	0.542**	1.00	
General Health	0.609**	0.066	0.506**	0.640**	-0.565**	0.513**	0.488**	0.395**	0.678**	0.526**	0.510**	0.429**	1.00

* Significant at the 5% level, **Significant at the 1% level. The Modified Rankin Scale is not included in this table as it is a 5 point categorical variable and therefore not appropriate for correlation analysis.

7.3.4. Physical Health and ICECAP-O

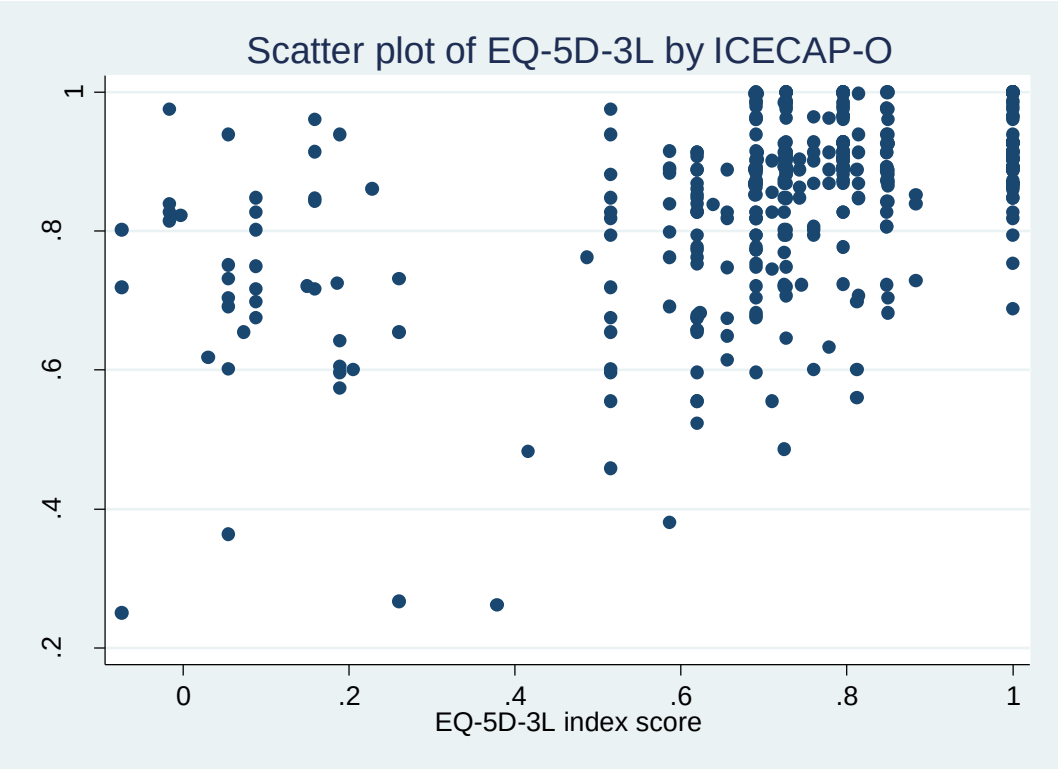
The baseline questionnaire pack administered to participants in the PastBP trial included a number of measures of physical health. These measures were the EQ-5D-3L, the physical sub-scales of the SF-36, the Modified Rankin Scale and the symptoms and side-effects questionnaire. This section presents the associations between the ICECAP-O and these measures.

7.3.4.1. The EQ-5D-3L

7.3.4.1.1. The ICECAP-O tariff and the EQ-5D-3L

Table 37 shows a moderate, statistically significant correlation of 0.518 between the ICECAP-O tariff score and the EQ-5D-3L index score. Figure 19 provides a graphical presentation of this correlation. This correlation indicates that as the EQ-5D-3L index scores increase, so do the ICECAP-o tariff scores. The statistical significance of this association means that it is unlikely to be due to chance.

Figure 19: Scatter plot of association between EQ-5D-3L index and ICECAP-O tariff



This result is in line with the correlation hypothesised *a priori*. It was expected that determinants other than health would also have a significant impact upon the person's capability as measured by the ICECAP-O. As noted previously, cross-sectional correlations are unable to confirm causality, but results indicate that while health is a notable determinant of a person's ICECAP-O scores, other factors may also influence it.

To assess whether this correlation held at different levels of health, correlations were assessed for the top quartile, the middle two quartiles and the bottom quartiles of EQ-5D-3L scores. Table 38 shows weak correlations at the higher EQ-5D-3L scores and negligible correlations at the lower levels of EQ-5D-3L scores.

Table 38: Correlations between the ICECAP-O tariff and EQ-5D-3L index at differing levels of health (n=446)

	EQ-5D-3L ≥0.85	EQ-5D-3L <0.85 & ≥0.69	EQ-5D-3L <0.62
ICECAP-O tariff	0.219*	0.202**	0.074

* Significant at the 5% level, **Significant at the 1% level.

Table 39 and Table 40 should be considered together. They provide information on the association of the ICECAP-O tariff score with the items of the EQ-5D-3L. In line with *a priori* hypotheses, the ICECAP-O tariff was associated with each of the EQ-5D-3L items at the 1% significance level.

Table 39: Associations between ICECAP-A tariff and EQ-5D-3L items (n=446)

Comparator	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
ICECAP-O tariff	**<0.001****	**<0.001****	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

Table 40 shows that for each of the EQ-5D-3L items, ICECAP-O tariff scores were higher in participants with better levels of health and lower in participants in worse health states. There were statistically significant differences between the ICECAP-O tariff scores for the top and second level for each EQ-5D-3L item. For Usual Activities and Pain/Discomfort there was a statistically significant difference between the bottom two levels. These differences were not seen in the three other EQ-5D-3L items. However, this could be related to the reduced statistical power at the lower levels due to small numbers. Table 40 provides strong evidence that in a medium sized, sample of older people, the ICECAP-O tariff scores are different across levels of the health dimensions in the EQ-5D-3L.

Table 40: Mean ICECAP-O tariff score by EQ-5D-3L item levels (n=447)

EQ-5D-3L Attribute	ICECAP-0 tariff score (95% CI)
Mobility	
No problems (n=223)	0.90 (089, 0.92)†
Some problems (n=223)	0.81 (0.79, 0.83)
Bed bound (n=1)	0.65
Self-care	
No problems (n=384)	0.88 (0.87, 0.89)†
Some problems (n=62)	0.74 (0.70, 0.78)
Unable (n=1)	0.72
Usual activities*	
No problems (n=260)	0.91 (0.89, 0.92)†
Some problems (n=168)	0.80 (0.78, 0.82)††
Unable (n=18)	0.65 (0.56, 0.74)
Pain and discomfort*	
No pain and discomfort (n=163)	0.89 (0.87, 0.91)†
Moderate pain and discomfort (n=254)	0.85 (0.83, 0.86)††
Extreme pain and discomfort (n=29)	0.76 (0.69, 0.82)
Anxiety and depression*	
Not anxious or depressed (n=320)	0.89 (0.87, 0.90)†
Moderately anxious or depressed (n=123)	0.79 (0.76, 0.81)
Extremely anxious or depressed (n=3)	0.52

Statistically significant (P<0.01) differences in the mean EQ-5D-3L index score between †† some problems/extreme problems response options and † no problems/some problems. * Items sum to less than 447 due to missing values.

7.3.4.1.2. ICECAP-O items and the EQ-5D-3L

Table 41 shows the associations between ICECAP-O items and the EQ-5D-3L index score. Statistically significant associations were seen between all items and the EQ-5D-3L index score. This provides mixed support for the *a priori* hypotheses. Expected associations (highlighted in bold) were seen for Security, Role, Enjoyment and Control, while an unexpected association between Attachment and the EQ-5D-3L index score was also seen.

Table 41: Associations between EQ-5D-3L index and ICECAP-O items (n=446)

Comparator	Attachment	Security	Role	Enjoyment	Control
EQ-5D-3L index score (n=442)	<0.001**	**<0.001****	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level. Hypothesised associations are highlighted in bold font

Table 42 can be considered together with Table 41. These tables indicate that the EQ-5D-3L index score was strongly and positively associated with ICECAP-O items. The possible exception to this was Attachment, which while shown to be significantly associated in Table 41, does not show significant differences between the levels in Table 42. The EQ-5D-3L index scores are greater at the higher capability levels of each ICECAP-O item. There were significant differences between the scores of the top, and second and third levels, of each ICECAP-O item apart from Attachment. There are no statistically significant differences between the bottom two levels of any item. This may be a result of low statistical power at these levels due to low numbers.

Table 42: Mean EQ-5D-3L index scores by ICECAP-O item levels (n=447)

Attribute and Level	EQ-5D-3L index score (95% CI)
Attachment*	
I can have all of the love and friendship that I want (n=268)	0.75 (0.72, 0.78)
I can have a lot of the love and friendship that I want (n=129)	0.71 (0.67, 0.75)
I can have a little of the love and friendship that I want (n=46)	0.67 (0.62, 0.73)
I cannot have any of the love and friendship that I want (n=3)	0.34
Security	
I can think about the future without any concern (n=115)	0.81 (0.77, 0.85)††
I can think about the future with only a little concern (n=202)	0.76 (0.73, 0.79)†
I can only think about the future with some concern (n=103)	0.62 (0.57, 0.68)
I can only think about the future with a lot of concern (n=27)	0.49 (0.39, 0.64)
Role*	
I am able to do all of the things that make me feel valued (n=158)	0.84 (0.81, 0.87)††
I am able to do many of the things that make me feel valued (n=196)	0.73 (0.70, 0.76)†
I am able to do a few of the things that make me feel valued (n=81)	0.54 (0.48, 0.60)
I am unable to do any of the things that make me feel valued (n=10)	0.40 (0.21, 0.57)
Enjoyment	
I can have all of the enjoyment and pleasure that I want (n=137)	0.84 (0.81, 0.87)††
I can have a lot of the enjoyment and pleasure that I want (n=216)	0.75 (0.72, 0.78)†
I can have a little of the enjoyment and pleasure that I want (n=86)	0.53 (0.46, 0.59)
I cannot have any of the enjoyment and pleasure that I want (n=8)	0.34 (0.13, 0.53)
Control	
I am able to be completely independent (n=231)	0.82 (0.80, 0.85)††
I am able to be independent in many things (n=171)	0.67 (0.63, 0.71)†
I am able to be independent in a few things (n=41)	0.46 (0.39, 0.54)
I am unable to be at all independent (n=4)	0.22

Statistically significant (P<0.01) differences in the mean EQ-5D-3L index score between †† all (any)/a lot (many) response options and † a lot (many)/few (a little). *Items sum to less than 447 due to missing values.

Table 43 shows the associations between the ICECAP-O items and the EQ-5D-3L items. A number of associations were seen at the 1% level of statistical significance. Expected associations are highlights in bold. Of the 19 hypothesised associations all were significant at the 1% level of significance. Five unexpected associations were identified, suggesting a closer association between the ICECAP-O and the EQ-5D-3L than was expected a priori.

Table 43: Associations between the ICECAP-O and EQ-5D-3L items (n=446)

Comparator	Attachment	Security	Role	Enjoyment	Control
EQ-5D-3L					
Mobility^{++}	0.001**	<0.001**	**<0.001****	**<0.001****	**<0.001****
Self-care^{++}	0.291	**<0.001****	**<0.001****	**<0.001****	**<0.001****
Usual^{++}	<0.001**	<0.001**	**<0.001****	**<0.001****	**<0.001****
Pain^{++}	0.203	**0.003****	**<0.001****	**<0.001****	**<0.001****
Anxiety^{++}	<0.001**	**<0.001****	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

Evidence of the direction of the relationship between the ICECAP-A items and the EQ-5D-5L items is provided in Table 44. Goodman & Kruskal's gamma (rank correlation) is used. All correlations are negative due to the difference in scoring between the ICECAP-A and the EQ-5D-3L (i.e. the top level on ICECAP items is 4, whereas the top level for EQ-5D-3L items is 1). Therefore, the results indicate that the better the health attribute score (e.g. greater mobility) the higher capability attribute score (e.g. more Control).

The associations which were expected a priori to be particularly strong are highlighted in bold. As can be seen, all these correlations were over 0.6, with the exception of the Enjoyment and pain/discomfort correlation which was unexpectedly low. Strong correlations were also seen between Enjoyment and usual activities and between Role and mobility.

Table 44: Correlations between ICECAP-O and EQ-5D-3L items (n=446)

Comparator	Attachment	Security	Role	Enjoyment	Control
EQ-5D-3L					
Mobility	-0.245	-0.365	-0.599	-0.513	**-0.689**
Self-care	-0.176	-0.478	-0.588	-0.568	**-0.817**
Usual activities	-0.325	-0.455	**-0.721**	-0.664	-0.771
Pain/discomfort	-0.117	-0.260	-0.436	**-0.423**	-0.444
Anxiety/depression	-0.293	-0.553	-0.573	**-0.616**	-0.564

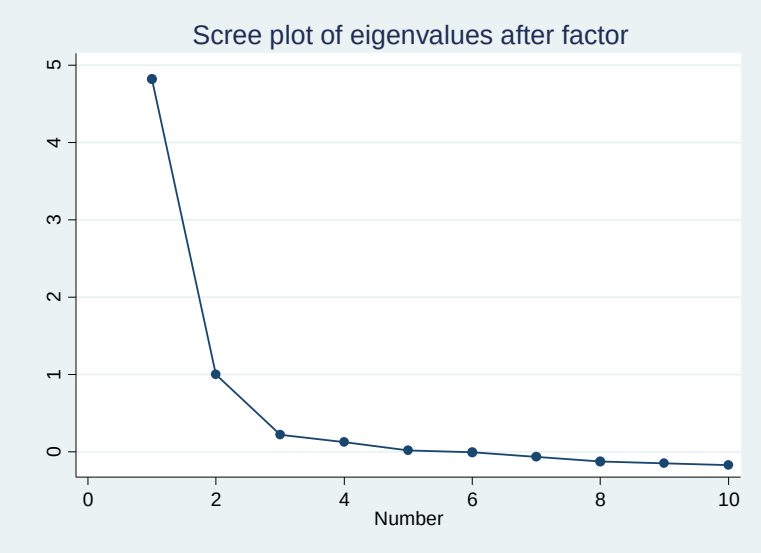
Highlighted in bold are the associations which were expected to be particularly strong.

7.3.4.1.3. Exploratory factor analysis

As with the ICECAP-A results from the BEEP trial, the results from the ICECAP-O in the PastBP trial indicate the possibility of a closer relationship with the EQ-5D-3L measure than previously anticipated. As described above, exploratory factor analysis provides a useful tool for further investigation of this relationship.

The number of factors to be retained was again judged on the Kaiser criterion which suggested that two factors was the optimal solution (see Figure 20).

Figure 20: Factor analysis eigen values by factor number



To control for the potential that the factors are correlated a Promax rotations was used. In a post-hoc test this was judged as the correct rotation as factors were shown to be correlated (-0.515) at a levels greater than 0.32, which is the cut off at which a oblique rotation becomes appropriate

Table 45 shows a two factor solution, indicating that the pooled items of the EQ-5D-3L and the ICECAP-O measure two different attributes. Results suggest that the ICECAP-A and

EQ-5D-3L measure two different underlying constructs. With the exception of anxiety/depression, all EQ-5D-3L items loaded onto factor one. Anxiety/depression and all ICECAP-O items, apart from control, loaded strongly onto factor 2. Control was split between factor one and two, with a somewhat greater loading onto factor one.

Table 45: Exploratory factor analysis comparing the ICECAP-A and EQ-5D-3L items (n=446)

	Rotated item loading on a 2-factor solution.	
	Factor 1	Factor 2
EQ-5D-3L		
Mobility	0.862	
Self-Care	0.818	
Usual Activities	0.819	
Pain	0.636	
Anxiety and Depression		-0.576
ICECAP-A		
Attachment		0.633
Security		0.678
Role	-0.218	0.703
Enjoyment		0.778
Control	-0.553	0.312
Factor correlations	-0.531	

Loadings of less than 0.2 are left blank to assist the interpretation of the results.

7.3.4.2. The SF-36

7.3.4.2.1. ICECAP-O tariff score and SF-36

The correlations between the ICECAP-O tariff score and the SF-36 subscales are presented in Table 46 [6]. All correlations were moderate and statistically significant at the 1% level. *A priori* expectations were that all SF-36 sub-scales would correlate moderately with the ICECAP-O tariff score. For the purposes of comparison, the correlations between the EQ-5D-3L and the SF-36 sub-scales are provided. Correlations between the ICECAP-O and the SF-36 sub-scales that were expected to be particularly strong are highlighted in bold. These correlations were all over 0.5. The mental health sub-scale and ICECAP-O tariff score also produced a correlation over 0.5. A comparison with the EQ-5D-3L shows that for the physical subscales there was a trend for stronger correlations with the EQ-5D-3L index score than the ICECAP-O tariff score for all subscales apart from general health. For the psychological sub-scales, correlations were similar on all scales except for mental health, where the correlation with the ICECAP-O tariff scores was notably stronger than with the EQ-5D-3L index.

[6] SF-36 T-scores were used in this analysis. As described in Table 6 a T-score has the advantage of standardising scores against a population norm which is fixed at 50. This is allows easier interpretation of scale scores (e.g. below 50 on any scale is a worse health state than the population norm).

Table 46: Correlations between ICECAP-O tariff and SF-36 sub-scales

SF-36	Correlation with ICECAP-O tariff	Correlation with EQ-5D-3L index
Physical		
Physical functioning (n=418)	0.467**	0.685**
Physical role (n=470)	0.435**	0.557**
Bodily Pain (n=474)	0.328**	0.696**
General Health (n=454)	**0.609****	0.506**
Psychological		
Vitality (n=457)	**0.599****	0.604**
Social functioning (n=478)	**0.527****	0.589**
Emotional role (n=462)	0.436**	0.399**
Mental health (n=472)	0.521**	0.382**

* Significant at the 5% level, **Significant at the 1% level. All scales were expected to be associated with the ICECAP-O tariff - highlighted in bold are the associations which were expected to be particularly strong.

7.3.4.2.2. ICECAP-O items and the SF-36

Table 47 shows the associations between the eight SF-36 sub-scales and the five ICECAP-O items. Of the 40 pairs, 39 showed significant associations at the 5% statistical significance level, while 38 of these associations were significant at the 1% level. Of the 31 hypothesised associations all were significant at the 1% level. Therefore, 8 associations were found that were not expected *a priori*. These findings provided mixed support for the *a priori* hypotheses: all hypothesised associations were significant, but a number of additional unexpected associations between the SF-36 and the ICECAP-O items were identified. These suggest a closer association between the SF-36 and the ICECAP-O than was anticipated.

Table 47: Associations between ICECAP-O items and SF-36 sub-scales

SF-36 scale	Attachment	Security	Role	Enjoyment	Control
Physical					
Phys func (n=418)	0.009**	**<0.001****	**<0.001****	**<0.001****	**<0.001****
Phys role (n=470)	0.012*	**<0.001****	**<0.001****	**<0.001****	**<0.001****
Pain (n=474)	0.400	**<0.001****	**<0.001****	**<0.001****	**<0.001****
Gen health (n=454)	<0.001**	**<0.001****	**<0.001****	**<0.001****	**<0.001****
Psychological					
Vitality (n=457)	<0.001**	<0.001**	**<0.001****	**<0.001****	**<0.001****
Social func (n=478)	**<0.001****	<0.001**	**<0.001****	**<0.001****	**<0.001****
Emo role (n=462)	**0.009****	<0.001**	**<0.001****	**<0.001****	**<0.001****
Mental health (n=472)	<0.001**	**<0.001****	**<0.001****	**<0.001****	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

The three SF-36 sub-scales demonstrating the strongest correlation with the ICECAP-O (General Health, Vitality and Social Functioning) were analysed in greater depth (below). Table 47 shows that these three subscales were significantly associated with all ICECAP-O items at the 1% significance level.

Table 48 indicates that SF-36 general health subscale scores are higher at the higher capability levels for each ICECAP-O item. Significant differences were found between the scores on the second and third levels for all items. On four of the items significant differences in the SF-36 general health subscale scores were found between the lower two levels, while a significant difference between the top two levels was only found on one item.

Table 48: Mean SF-36 general health sub-scale score by ICECAP-O item level (n=421)

Attribute and Level	SF-36 General Health sub-scale t score (95% CI)
Attachment	
I can have all of the love and friendship that I want (n=249)	46.48 (45.24, 47.73)
I can have a lot of the love and friendship that I want (n=122)	44.30 (42.73, 45.87)††
I can have a little of the love and friendship that I want (n=46)	39.54 (36.86, 42.23)
I cannot have any of the love and friendship that I want (n=3)	35.29 (-3.52, 74.11)
Security	
I can think about the future without any concern (n=107)	51.42 (50.06, 52.78)†††
I can think about the future with only a little concern (n=191)	45.93 (44.74, 47.11)††
I can only think about the future with some concern (n=95)	39.26 (37.29, 41.11)†
I can only think about the future with a lot of concern (n=28)	33.85 (30.46, 37.23)
Role	
I am able to do all of the things that make me feel valued (n=147)	50.97 (49.68, 52.26)
I am able to do many of the things that make me feel valued (n=186)	43.97 (42.84, 45, 10)††
I am able to do a few of the things that make me feel valued (n=79)	37.77 (35.56, 39.97)†
I am unable to do any of the things that make me feel valued (n=8)	30.23 (20.86, 39.60)
Enjoyment	
I can have all of the enjoyment and pleasure that I want (n=127)	51.19 (49.81, 52.56)
I can have a lot of the enjoyment and pleasure that I want (n=203)	45.13 (44.06, 46.21)††
I can have a little of the enjoyment and pleasure that I want (n=85)	36.56 (34.49, 38.62)†
I cannot have any of the enjoyment and pleasure that I want (n=6)	30.13 (19.01, 41.25)
Control	
I am able to be completely independent (n=217)	49.01 (47.90, 50.12)
I am able to be independent in many things (n=160)	42.50 (41.18, 43.83)††
I am able to be independent in a few things (n=39)	34.87 (31.55, 38.19)†
I am unable to be at all independent (n=5)	31.00 (15.01, 46.99)

Statistically significant differences in the mean SF-36 general health sub-scale score between † unable and a little, †† a little and a lot and ††† a lot and all of the things.

Table 49 shows higher scores on the SF-36 vitality sub-scale were seen at higher capability levels on each of the ICECAP-O items. Pronounced differences in scores between the top two levels and the second and third levels on each item were found. Differences in scores between the bottom two capability levels were less pronounced.

Table 49: Mean SF-36 vitality sub-scale score by ICECAP-O item level (n=457)

Attribute and Level	SF-36 Vitality sub-scale t score (95% CI)
Attachment	
I can have all of the love and friendship that I want (n=256)	50.23 (48.94, 51.52)
I can have a lot of the love and friendship that I want (n=120)	48.93 (47.16, 50.69)††
I can have a little of the love and friendship that I want (n=44)	41.70 (38.97, 44.44)
I cannot have any of the love and friendship that I want (n=3)	40.85 (-3.88, 85.59)
Security	
I can think about the future without any concern (n=111)	55.23 (53.61, 56.85) †††
I can think about the future with only a little concern (n=184)	49.77 (48.37, 51.16)††
I can only think about the future with some concern (n=101)	43.20 (41.35, 45.06)†
I can only think about the future with a lot of concern (n=28)	39.07 (35.56, 42.57)
Role	
I am able to do all of the things that make me feel valued (n=152)	54.98 (53.66, 56.30) †††
I am able to do many of the things that make me feel valued (n=185)	48.26 (46.94, 49.59)††
I am able to do a few of the things that make me feel valued (n=77)	39.85 (37.62, 42.07)
I am unable to do any of the things that make me feel valued (n=9)	37.52 (30.87, 44.17)
Enjoyment	
I can have all of the enjoyment and pleasure that I want (n=131)	54.92 (53.41, 56.43) †††
I can have a lot of the enjoyment and pleasure that I want (n=199)	49.55 (48.35, 50.75)††
I can have a little of the enjoyment and pleasure that I want (n=87)	39.27 (37.20, 41.35)
I cannot have any of the enjoyment and pleasure that I want (n=7)	39.07 (32.05, 46.09)
Control	
I am able to be completely independent (n=221)	52.99 (51.76, 54.22) †††
I am able to be independent in many things (n=160)	46.22 (44.77, 47.68)††
I am able to be independent in a few things (n=38)	38.68 (35.77, 41.60)
I am unable to be at all independent (n=5)	33.86 (15.59, 52.13)

Statistically significant differences in the mean SF-36 vitality sub-scale score between † unable and a little, †† a little and a lot and ††† a lot and all of the things.

Table 50 indicates that on all ICECAP-O items the higher capability levels, the higher the scores on the SF-36 social functioning sub-scale. On the items of Role and Enjoyment, which might be considered to be the closest of the ICECAP items to social functioning, significant and pronounced differences were seen between the bottom two levels: those who could not have any enjoyment or were unable to do any of the things that made them feel valued, scored very low on the social functioning sub-scale.

Table 50: Mean SF-36 social functioning sub-scale scores by ICECAP-O item level (n=478)

Attribute and Level	SF-36 social functioning sub-scale t score (95% CI)
Attachment	
I can have all of the love and friendship that I want (n=263)	49.07 (47.79, 50.35)
I can have a lot of the love and friendship that I want (n=128)	46.88 (44.92, 48.84)††
I can have a little of the love and friendship that I want (n=49)	41.71 (38.61, 44.82)
I cannot have any of the love and friendship that I want (n=3)	35.03 (-13.81, 83.87)
Security	
I can think about the future without any concern (n=113)	53.33 (52.08, 54.57)†††
I can think about the future with only a little concern (n=200)	48.59 (47.21, 49.96)††
I can only think about the future with some concern (n=103)	42.45 (40.05, 44.84)†
I can only think about the future with a lot of concern (n=28)	35.62 (30.86, 40.38)
Role	
I am able to do all of the things that make me feel valued (n=154)	52.78 (51.60, 53.96)
I am able to do many of the things that make me feel valued (n=196)	48.08 (46.68, 49.48)††
I am able to do a few of the things that make me feel valued (n=82)	39.02 (36.44, 41.61)†
I am unable to do any of the things that make me feel valued (n=10)	29.58 (20.38, 38.77)
Enjoyment	
I can have all of the enjoyment and pleasure that I want (n=134)	52.49 (51.05, 53.93)
I can have a lot of the enjoyment and pleasure that I want (n=213)	48.40 (47.09, 49.71)††
I can have a little of the enjoyment and pleasure that I want (n=89)	39.94 (37.48, 42.39)†
I cannot have any of the enjoyment and pleasure that I want (n=8)	26.85 (17.41, 36.29)
Control	
I am able to be completely independent (n=229)	51.13 (49.96, 52.31)
I am able to be independent in many things (n=169)	45.20 (43.53, 46.87)††
I am able to be independent in a few things (n=41)	38.89 (35.12, 42.67)†
I am unable to be at all independent (n=5)	33.94 (7.54, 60.34)

Statistically significant differences in the mean SF-36 social functioning sub-scale score between † unable and a little, †† a little and a lot and ††† a lot and all of the things.

7.4.3.3. Modified Ranking Scale

The Modified Rankin Scale scores and the ICECAP-O tariff scores were significantly associated at the 1% level, which is in line with *a priori* hypotheses. Table 51 shows a trend of higher ICECAP-O tariff scores at lower levels of disability. Differences in ICECAP-O tariff scores are more pronounced at the bottom levels of the Modified Rankin Scale, which indicate moderate and moderately severe disability.

Table 51: Mean ICECAP-O tariff score by Modified Rankin Scale score (n=452)

Modified Ranking Scale score	ICECAP-O tariff score
Level 0	0.92 (0.9, 0.94)
Level 1	0.89 (0.88, 0.91)††
Level 2	0.84 (0.83, 0.86)†
Level 3	0.76 (0.72, 0.79)
Level 4	0.69 (0.60, 0.78)

Statistically significant differences in the mean SF-36 social functioning sub-scale score between †† no significant disability and slight disability and †slight disability and moderate disability.

An item-by-item analysis, presented in Table 52, shows that the Modified Ranking Scale scores were significantly associated with each ICECAP-O item. Two additional associations (with Attachment and Enjoyment) were found that were not hypothesised. This suggests a closer than expected association between the ICECAP-O and the Modified Rankin Scale.

Table 52: Associations between Modified Rankin Scale and ICECAP-O items (n=452)

Comparator	Attachment	Security	Role	Enjoyment	Control
Modified Rankin Scale	<0.001**	**<0.001****	**<0.001****	<0.001**	**<0.001****

* Significant at the 5% level, **Significant at the 1% level.

7.3.4.5. Symptoms and side-effects

Table 37 shows a moderate, statistically significant correlation of -0.477 between the ICECAP-O tariff score and the number of symptoms and side effects a person reports. This correlation indicates that the higher the number of symptoms and side-effects a person

reports, the lower the ICECAP-O tariff score will be. The symptoms and side-effects analysis was an exploratory analysis.

To assess the association further, symptoms and side effects were transformed into a categorical variable of: few (3 or less reported symptoms or side-effects), some (more than 3 and 9 or less) and many (more than 9). Table 53 shows the ICECAP-O tariff scores by group for those who have few, some and many side-effects. The group with few side effects had higher mean ICECAP-O tariff scores than those with some, and those with some had higher scores than those with many. The differences between the groups were statistically significant at the 1% level.

Table 53: Mean ICECAP-O tariff scores by number of symptoms and side-effects (n=299)

	ICECAP-O tariff score (95% CI)
Few (Equal to or less than 3)	0.92 (0.90, 0.94)††
Some (More than 3 and less than or equal to 9)	0.85 (0.83, 0.87)†
Many (More than 9)	0.75 (0.71, 0.79)

Statistically significant difference between †† Few/Some and † Some/Many symptoms and side-effects at P<0.01 level.

Table 54 shows the associations between the categorical symptoms and side-effects variables and each of the ICECAP-O items. All of the ICECAP-O items were significantly associated with the number of symptoms and side-effects at the 1% significance level.

Table 54: Associations between number symptoms and side-effects and ICECAP-O items (n=299)

	Attachment	Security	Role	Enjoyment	Control
SSE	0.001**	<0.001**	<0.001**	<0.001**	<0.001**

* Significant at the 5% level, **Significant at the 1% level.

7.3.5. Comparison of results with hypotheses

In line with the hypothesis based approach to validity testing, hypotheses were formed *a priori*. Table 55 presents a comparison of hypotheses and the results of the PastBP trial analysis. These results provide positive, supportive evidence of ICECAP-O validity in the PastBP trial patient population. The overwhelming majority of hypothesised associations were confirmed by the results. Table 55 shows confirmatory evidence of the hypothesised associations between the ICECAP-O tariff score and the socio-demographic, physical health and psychological health variables. The strength of the correlations were moderate to weak, which was as expected.

Table 55: A comparison of hypotheses and results from the PastBP trial

Hypotheses				Results			
	Association	Direction	Strength		Association	Direction	Strength
Socio-demographic				**Socio-demographic**			
Age	Yes			Age	No		
Gender	No			Gender	No		
Physical health				**Physical health**			
EQ-5D-3L	Yes	Positive		EQ-5D-3L	Yes	Positive	Moderate
EQ-5D-3L VAS	Yes	Positive		EQ-5D-3L VAS	Yes	Positive	Moderate
SSE	*Exploratory analysis*			SSE	Yes	Negative	Moderate
Physical functioning	Yes	Positive		Physical functioning	Yes	Positive	Moderate
Physical role	Yes	Positive		Physical role	Yes	Positive	Moderate
Bodily pain	Yes	Negative		Bodily pain	Yes	Negative	Weak
General heatlh	Yes	Positive		General heatlh	Yes	Positive	Moderate
Psychological health				**Psychological health**			
Vitality	Yes	Positive		Vitality	Yes	Positive	Moderate
Social functioning	Yes	Positive		Social functioning	Yes	Positive	Moderate
Emotional role	Yes	Positive		Emotional role	Yes	Positive	Moderate
Mental health	Yes	Positive		Mental health	Yes	Positive	Moderate

No significant association between gender and ICECAP-O scores was expected: none was seen. An association between age and the ICECAP-O tariff score was expected but not found. Item by item analysis provided mixed support for the hypotheses: as hypothesised the ICECAP-O attribute of Control, was significantly associated with age, and an unexpected association was found between age and Security.

The associations between the measures of physical and psychological health in the PastBP trial and the ICECAP-A tariff score were all as expected. The EQ-5D-3L index score and the EQ-5D-3L VAS score both showed the moderate, positive correlation which was hypothesised *a priori*. All the subscales of the SF-36 showed the expected direction and strength of correlation.

The item-by-item analysis of the ICECAP-O showed some notable deviations from the *a priori* hypotheses. Where an association was expected, these associations were found in the anticipated direction. However, for both the EQ-5D-3L and the SF-36 sub-scales there were a number of associations found with the ICECAP-O items that were not expected *a priori*. These "extra" associations are suggestive of a closer relationship between the ICECAP-O and measures of physical and psychological functioning than anticipated. The mobility and usual activity items showed additional associations with Security, while Attachment showed additional associations with the usual activities and anxiety/depression items. The four psychological sub-scales of the SF-36 were unexpectedly correlated with the Security item of the ICECAP-O. The physical sub-scales of the SF-36 were unexpectedly associated with Attachment, with 3 unexpected associations between pairs.

CHAPTER 8. QUANTITATIVE STUDY OF THE VALIDITY AND RESPONSIVENESS OF THE ICECAP MEASURES: RESPONSIVENESS RESULTS

8.1. Chapter introduction

This chapter reports the results of two separate anchor based responsiveness analyses for the ICECAP-A and ICECAP-O measures using data from the BEEP and PastBP trials respectively. The results from the ICECAP-A analysis are reported first, followed by the ICECAP-O analysis.

A description of the participants included in each analysis is provided at the start of each section. The choice of anchors is then described. Multiple anchor analyses are then presented for each measure. Each anchor analysis consists of: 1) an item-by-item analysis of the change in response profiles of the ICECAP measures; 2) an inspection of the correlation between the non-weighted ICECAP scores and the anchor and an analysis of the change in non-weighted ICECAP scores in participants whose anchor scores have improved or worsened; 3) an inspection of the correlation between the ICECAP tariff scores and the anchor and an analysis of the change in ICECAP tariff scores in participants whose anchor scores have improved or worsened.

As described in Chapter 3, it is recommended that numerous anchors are used in a responsiveness analysis [261]. This has the practical implication of making the presentation of results repetitive. To reduce repetition, a number of the analyses have been placed in appendices and summaries provided in this chapter. The analyses retained in full were chosen on the basis that, together, they provide a broad and varied selection of anchors measuring general health, psychological health, disability and social functioning. The EQ-5D-3L has been retained in both ICECAP-A and ICECAP-O analyses due to the frequency with which it is currently used in health economic analysis. The choice of those anchors

retained in the chapter was not based on whether the results provide positive or negative indications of ICECAP measure responsiveness.

8.2. BEEP trial

8.2.1. Participant characteristics

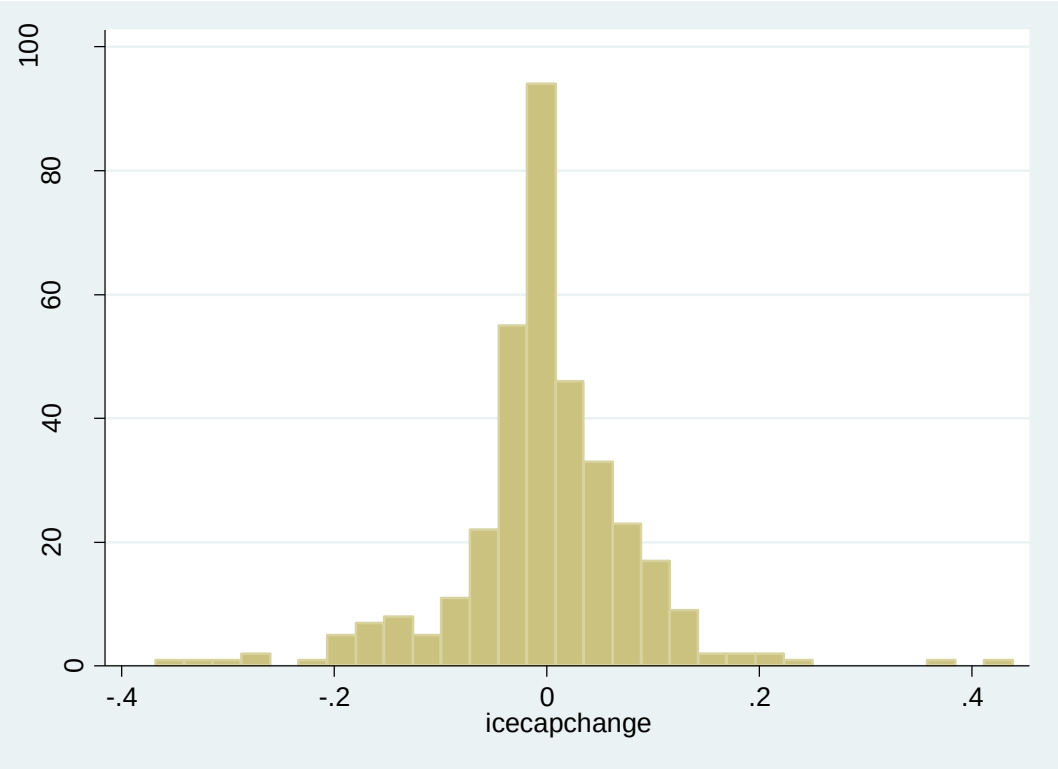
The characteristics of the BEEP trial participants used in this responsiveness analysis are presented in Table 56. The average age of participants is 64, with a roughly equal proportion of male and female participants. The average ICECAP-A capability tariff scores were higher at both baseline and follow-up than values reported in the general population [116]. Participants reported a mean baseline value of 0.64 for the EQ-5D-3L, which was below the UK national average for this age group [311]. The mean GAD-7 and PHQ-8 scores did not indicate wide-spread anxiety disorder or depression within the sample [295,296]. The co-morbidities questionnaire completed at baseline, was not completed at 6 month follow-up; therefore change in health could not be assessed using this questionnaire.

Table 56 shows that the mean value of the ICECAP-A tariff has not changed between baseline and follow-up. Mean change can "hide" individual change and when completing a responsiveness analysis the range of change that is present in a sample is an important consideration. Figure 21 shows the distribution of change amongst BEEP participants. As can be seen, the majority of participants either do not change or change by less than 0.1. Therefore, this responsiveness analysis was completed in a population that showed small changes in capability and had high levels of capability at baseline.

Table 56: Characteristics BEEP trial participants, including mean scores and median scores

Attribute	Mean baseline values (SD)	Median baseline value (IQR)	Mean follow-up value (SD)	Median follow-up value (IQR)	Measure range	Sample size
Socio-demographic						
Age mean	63.9 (9.83)					357
Gender (% male)	49.3%					357
Health and functioning						
ICECAP-O tariff	0.89 (0.11)	0.91 (0.85, 0.97)	0.89 (0.12)	0.92 (0.85, 0.97)	0 to 1	355
EQ-5D-3L index	0.64 (0.23)	0.69 (0.62, 0.76)	0.70 (0.22)	0.73 (0.69, 0.8)	-0.59 to 1	351
WOMAC Pain	8.30 (3.45)	8 (6, 11)	6.25 (3.81)	6 (3, 9)	0 to 18	353
WOMAC Stiffness	3.66 (1.72)	4 (3,5)	2.98 (1.77)	3 (2, 4)	0 to 8	357
WOMAC Functioning	27.63 (12.1)	27 (20, 37)	21.55 (13.68)	20 (10, 32)	0 to 62	353
GAD-7	3.14 (4.41)	1 (0, 4)	2.50 (3.91)	1 (0, 4)	0 to 21	344
PHQ-8	3.69 (4.44)	2 (1, 5)	2.99 (3.89)	2 (0, 4)	0 to 24	341

Figure 21: Frequency distribution of change in ICECAP-A tariff score



8.2.2. Choice of anchors from BEEP trial

Table 57 shows the correlation of change scores between measures administered at baseline and follow-up. The strongest correlations were seen between the WOMAC subscales, between the GAD-7 and PHQ-8 and between the WOMAC subscale of pain and EQ-5D-3L. All other correlations were below 0.3, with the majority of correlations with the ICECAP-A tariff below 0.2. The strongest correlations were with the EQ-5D-3L index score, GAD-7 and PHQ-8.

The process through which anchors were chosen for the responsiveness analysis is described in greater detail in Chapter 3 and Chapter 6. Based on the strength of correlations between ICECAP-A and other measures, and other considerations previously described, the following assessments of responsiveness will be completed:

- An assessment using the EQ-5D-3L
- An assessment using the GAD-7
- An assessment using the PHQ-8 (placed in appendices)

Table 57: Correlations between change scores of measures in the BEEP trial

	ICECAP-O tariff	EQ-5D-3L index	WOMAC Pain	WOMAC Stiffness	WOMAC Functioning	GAD-7	PHQ-8
ICECAP-A tariff	1.00						
EQ-5D-3L Index	0.255	1.00					
WOMAC Pain	-0.055	-0.402	1.00				
WOMAC Stiffness	-0.151	-0.236	0.507	1.00			
WOMAC Functioning	-0.103	-0.380	0.737	0.592	1.00		
GAD-7	-0.205	-0.202	0.109	0.040	0.129	1.00	
PHQ-8	-0.190	-0.232	0.057	0.092	0.090	0.511	1.00

8.2.3. EQ-5D-3L Index anchor analysis

8.2.3.1. Anchor group formation

EQ-5D-3L anchor groups were formed using the minimally important difference values taken from Walters and Brazier of 0.074 [262]. Change groups were formed of participants whose health status had improved or worsened by equal to or greater than +/-0.074 on the EQ-5D-3L index. The mean change in EQ-5D-3L index score in each change group was therefore larger than 0.074 (values included in Table 58); therefore these are not groups of minimally important change, rather groups of participants who all report at least a minimally important change[7]. EQ-5D-3L change was similar in the improved and worsened groups. These changes are shown in Table 58.

Table 58: Group numbers and mean EQ-5D-3L index change in the EQ-5D-3Lindex anchor groups (n=341)

Anchor group	Number in group	Mean EQ-5D-3L change in group (95% CI)	Mean EQ-5D 3L change as % of possible change
Improved	97	0.29 (0.254, 0.326)	18.2%
No change	206	0.01 (0.001, 0.013)	0.4%
Worsened	38	-0.31 (-0.373, -0.241)	19.3%

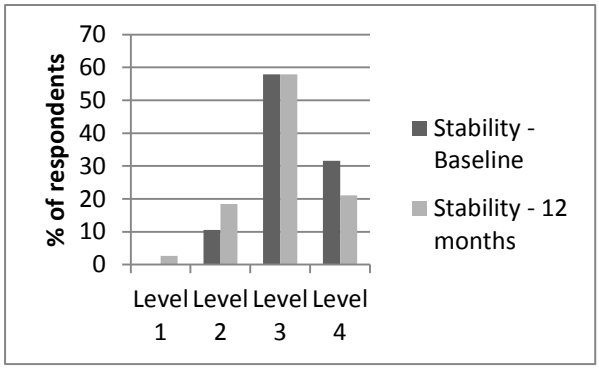
[7] The mean change in all anchor groups presented in this chapter is greater than the minimally important change used to form them. This can be assessed by the reader through considering the mean anchor change and the mean anchor change as a percentage of possible change presented at the start of each section.

8.2.3.2. Item by item analysis

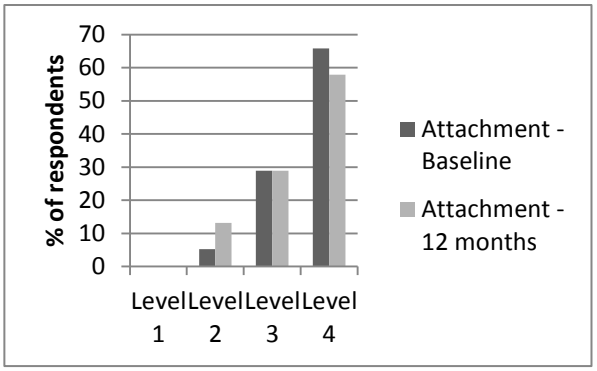
Figure 22 shows the response profiles at baseline and follow-up for the group of respondents who reported a reduction in their EQ-5D-3L index scores. These results are presented in numerical form in Appendix 23. The response profile of these participants changed, with reductions in percentage of respondents answering level 4 (full capability) of all items. The largest reductions in the percentage of respondents answering level 4 was for the Enjoyment and Autonomy item, where in both cases the reduction was 21 percentage points. Smaller reductions, of between 6 and 11 percentage points was found for the other items. For all items minimal change occurred in the percentage of respondents answering the bottom level of capability (level 1).

Figure 22: ICECAP-A response profile at baseline and follow-up for participants reporting a worsening of their EQ-5D-3L index scores

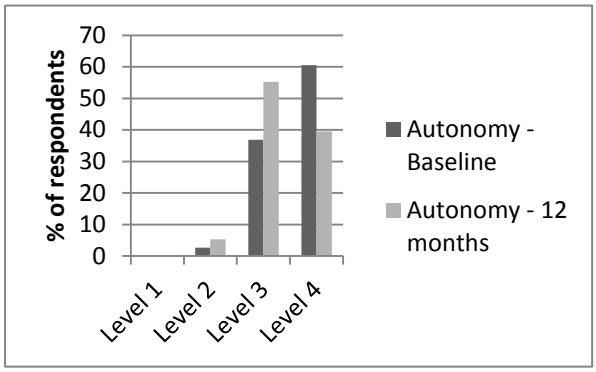
Stability Item



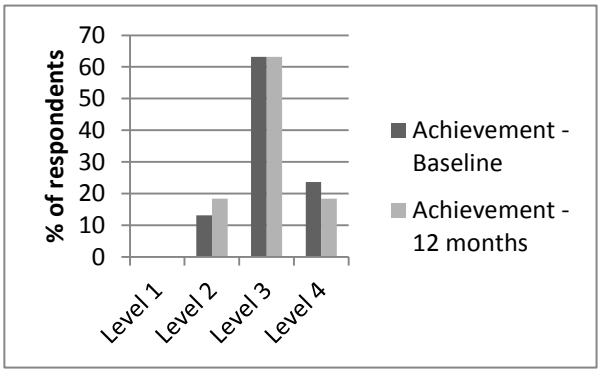
Attachment Item



Autonomy Item



Achievement Item



Enjoyment Item

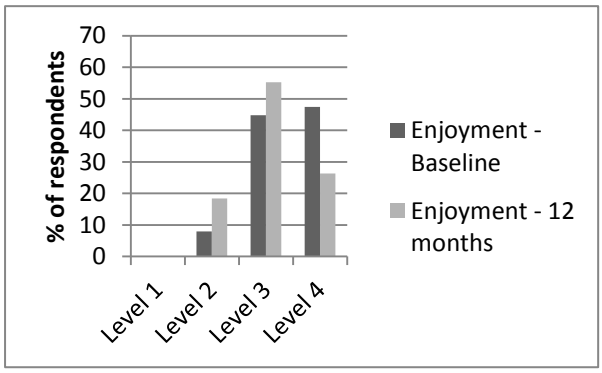
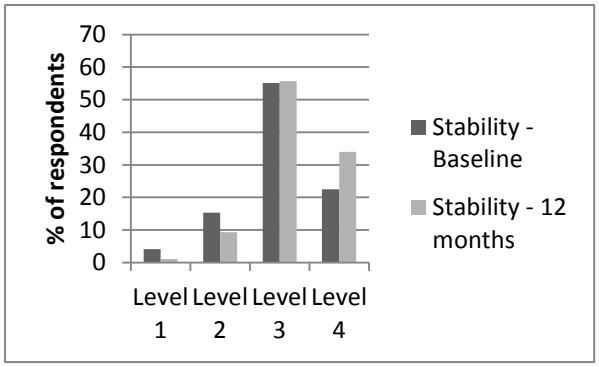


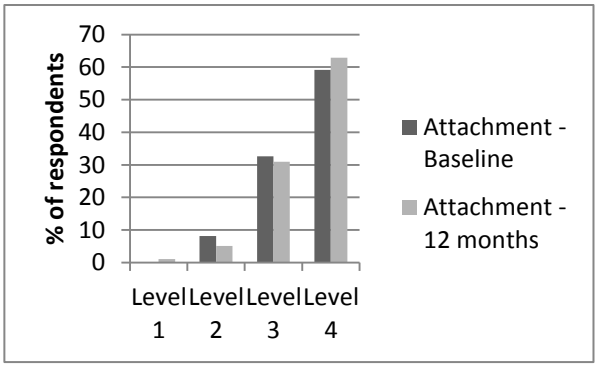
Figure 23 shows changes in the response profiles of the ICECAP-A in participants reporting an improvement in their EQ-5D-3L index scores. These results are presented in numerical form in Appendix 24. The change in the response profile of participants reporting an improvement in EQ-5D-3L scores was less pronounced than the change seen for participants who reported a worsening. Increases of 8, 13 and 10 points were seen in the percentage of participants reporting full-capability (level 4) on Stability, Autonomy and Achievement respectively. Smaller changes of 4 and 1 points in the percentage of respondents answering the top level of Attachment and Enjoyment were found. There was minimal change in the percentage of respondents answering the bottom level on each item.

Figure 23: ICECAP-A response profile at baseline and follow-up for participants reporting an improvement in their EQ-5D-3L index scores

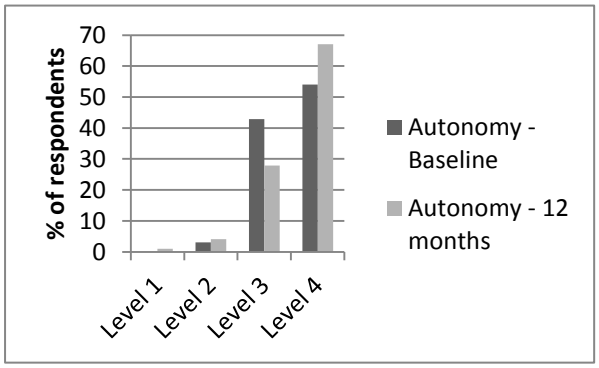
Stability Item



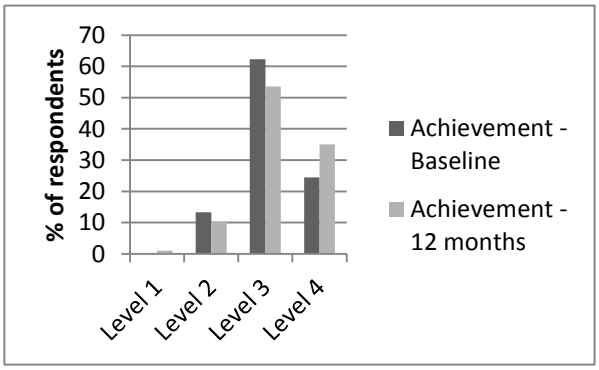
Attachment Item



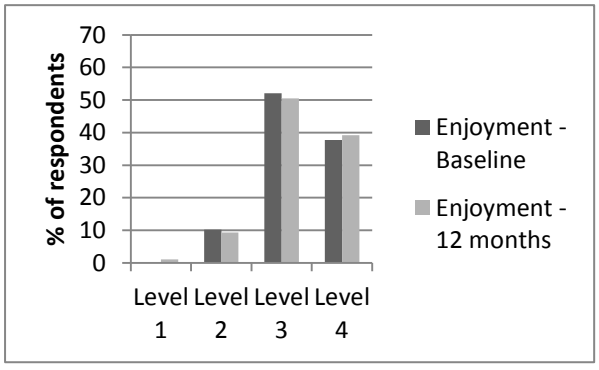
Autonomy Item



Achievement Item



Enjoyment Item



8.2.3.3. Non-weighted ICECAP scores

Table 59 shows that cross-sectional correlations between the EQ-5D-3L index and the non-weighted ICECAP-A score at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of the change in these measures between baseline and 6 month follow-up was weak and significant at the 1% level.

Table 59: Cross-sectional and change correlations between EQ-5D-3L index and non-weighted ICECAP-A scores (n=341)

	ICECAP-A		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
EQ-5D-3L index score	0.477**	0.512**	0.191**

* Significant at the 5% level, **Significant at the 1% level.

Table 60 shows change in non-weighted ICECAP-A scores by EQ-5D-3L index anchor change groups. In the group of participants reporting an improvement in EQ-5D-3L scores, the mean non-weighted ICECAP-A score increased. In the group reporting a worsening of their EQ-5D-3L index scores the mean non-weighted ICECAP-A score decreased. The change in ICECAP-A score was larger in the group that had worsened than in the group that had improved. The mean EQ-5D-3L index score change in these anchor groups was similar; therefore the proportion of ICECAP-A change to EQ-5D-3L anchor change was larger for the group that had worsened than for the group that had improved. The effect sizes for the improved group were small, while for the group that had worsened they were moderate (or approaching moderate).

Table 60: Mean change in non-weighted ICECAP-A scores by EQ-5D-3L index anchor change groups (n=341)

Anchor group	Baseline ICECAP-A scores	Follow-up ICECAP-A scores	Mean ICECAP-A change (95% CI)	Change as a % of possible change	ES	SRM
Improved	16.423	16.897	0.474** (-0.123, 0.826)	3.2%	0.2	0.27
No change	17.131	17.150	0.019 (-0.190, 0.229)	0.1%	0.01	0.01
Worsened	16.895	15.842	-1.053** (0.496, 1.609)	7.0%	0.47	0.62

* Significant at the 5% level, **Significant at the 1% level.

8.2.3.4. ICECAP-A tariff score

Table 61 shows the correlations between the ICECAP-A tariff and the EQ-5D-3L index at baseline, follow-up and over time. Baseline and follow-up cross-sectional correlations were moderate and significant at the 1% level. The correlation of the change scores between baseline and 6 month follow-up was weak and significant at the 1% level.

Table 61: Cross-sectional and change correlations between the EQ-5D-3L index and ICECAP-A tariff (n=341)

	ICECAP-A		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
EQ-5D-3L index	0.458**	0.484**	0.255**

* Significant at the 5% level, **Significant at the 1% level.

Table 62 and Figure 24 show change in ICECAP-A tariff score by EQ-5D-3L index anchor groups. In the group of participants reporting an improvement in the EQ-5D-3L scores, a small mean increase in ICECAP-A tariff scores was observed, which was significant at the 5% level. The effect sizes and SRMs for this change were small. In participants reporting a

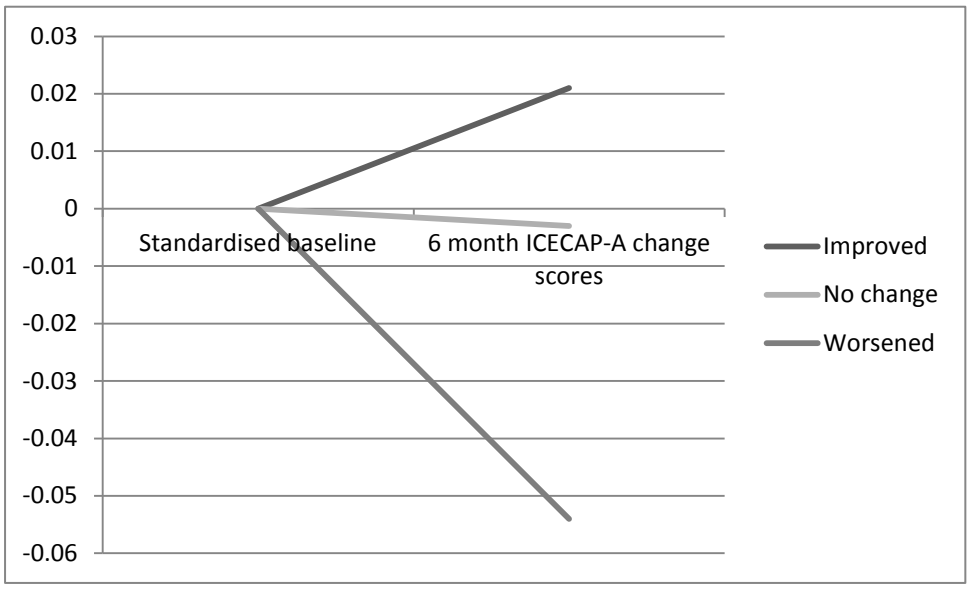
reduction in the EQ-5D-3L scores, a moderate reduction in ICECAP-A tariff scores was seen, which was significant at the 1% level. The effect sizes and SRMs were moderate. There are differences in the mean change of the ICECAP-A between patients who reported improved or reduced EQ-5D-3L scores, with those reporting a worsening reporting a larger reduction. Change as a percentage of possible change was smaller for the ICECAP-A tariff score than for the non-weighted ICECAP-A scores.

Table 62: Mean change in ICECAP-A tariff scores by EQ-5D-3L index anchor change groups (n=341)

Anchor group	Baseline ICECAP-A scores	Follow-up ICECAP-A scores	Mean ICECAP-A change (95% CI)	Change as a % of possible change	ES	SRM
Improved	0.863	0.884	0.021* (0.001, 0.041)	2.1%	0.17	0.21
No change	0.898	0.895	-0.003 (-0.128, 0.007)	0.3%	0.02	0.03
Worsened	0.890	0.836	-0.054** (-0.084, -0.024)	5.4%	0.53	0.59

* Significant at the 5% level, **Significant at the 1% level.

Figure 24: Mean change in ICECAP-A tariff scores by EQ-5D-3L index anchor change groups (n=341)



8.2.4. GAD-7 anchor analysis

8.2.4.1. Anchor group formation

Anchor groups from the GAD-7 score were not formed using a minimally important difference as no value could be found in the existing literature. The option of using the inter-quartile range of change was examined. The interquartile range values were -1 and 0. These values were not considered a large enough change for this analysis (e.g. using a change value of 0 to define an anchor group is not possible). Therefore, an arbitrary value of change by 2 points on the GAD-7 was used to define groups. This value allowed adequate numbers in each of the change groups (see Table 63). No assumption can be made as to whether this change was important to participants.

Table 63: Group numbers and mean GAD-7 change scores in GAD-7 anchor groups (n=335)

Anchor group	Number in group	Mean GAD-7 change in group (95% CI)	Mean GAD-7 change as % of possible change
Improved	83	4.843 (4.043, 5.642)	23.1%
No change	209	0.047 (-0.032, 0.126)	0.2%
Worsened	43	-4.93 (-6.076, -3.783)	23.5%

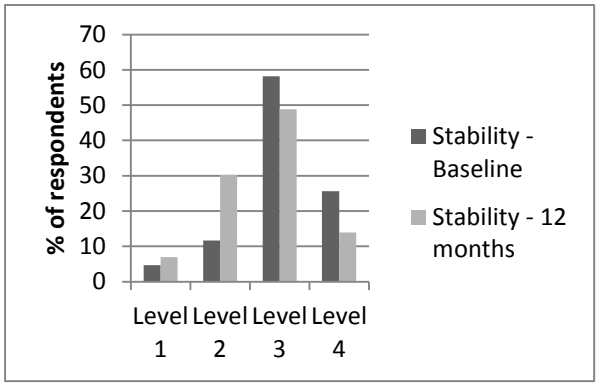
8.2.4.2. Item by item analysis

Figure 25 shows the response profiles at baseline and follow-up for the group of respondents reporting a worsening of their GAD-7 psychological health status (an increase in their GAD-7 scores). These results are presented in numerical form in Appendix 25. Reductions of 14

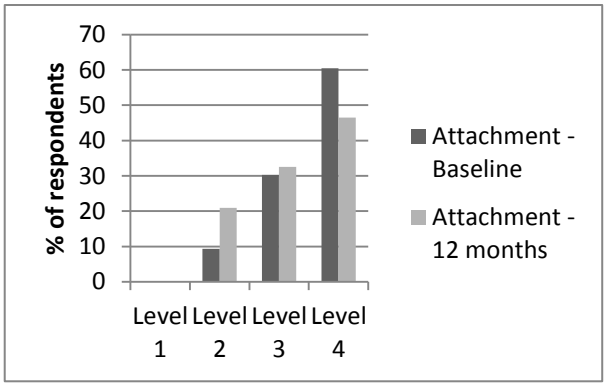
percentage points in respondents answering the level 4 (full capability) of the Attachment and Enjoyment were found. Reductions of 12 and 9 percentage points were found in those answering the top level of Stability and Achievement items respectively. Little change was found in the Autonomy item. There was notable change in the percentage of respondents answering the lower capability levels (level 1 and 2) for Stability and Attachment items.

Figure 25: ICECAP-A response profile at baseline and follow-up for participants reporting a worsening in their GAD-7 health status

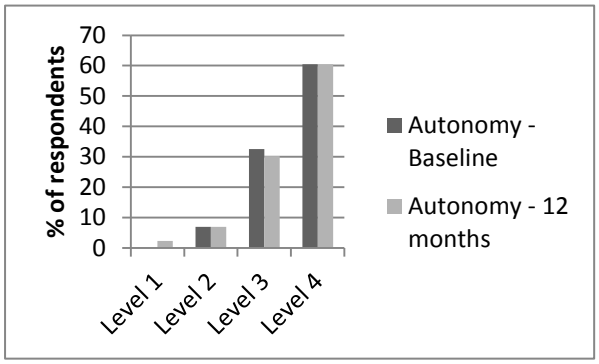
Stability Item



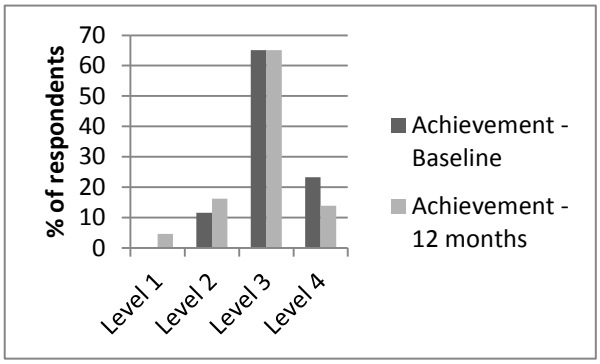
Attachment Item



Autonomy Item



Achievement Item



Enjoyment Item

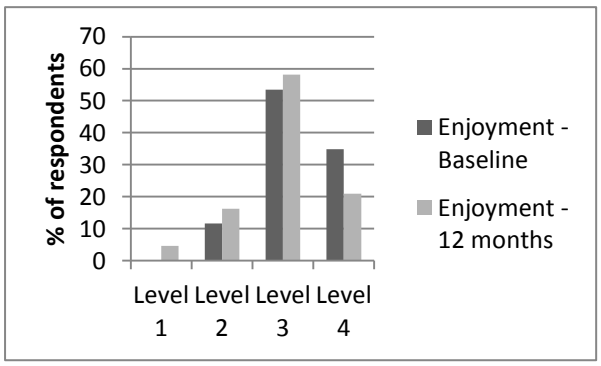
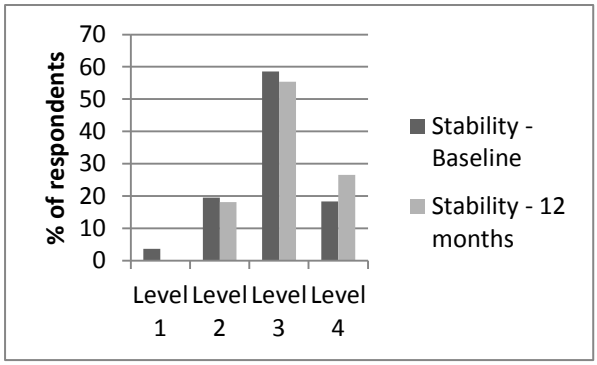


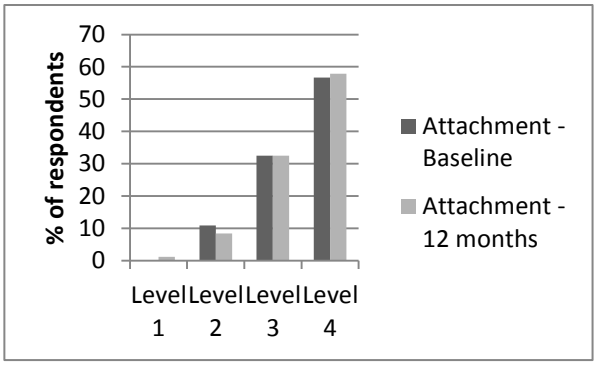
Figure 26 shows response profiles at baseline and follow up for the group of respondents who reported an improvement in their GAD-7 psychological health status (a reduction in GAD-7 scores). These results are presented in numerical form in Appendix 26. Changes were smaller and less pronounced than changes in the group reporting a worsening of their GAD-7 psychological health status. The largest changes in the percentage of respondents answering level 4 were found in the Stability and Enjoyment items where there were 9 and 6 percentage point increases respectively. Smaller changes were found in the other items. In comparison to the group of respondents reporting worsening GAD-7 psychological health status, less pronounced change in the percentage of respondents answering the bottom two levels was found.

Figure 26: ICECAP-A response profile at baseline and follow-up for participants reporting an improvement in their GAD-7 health status

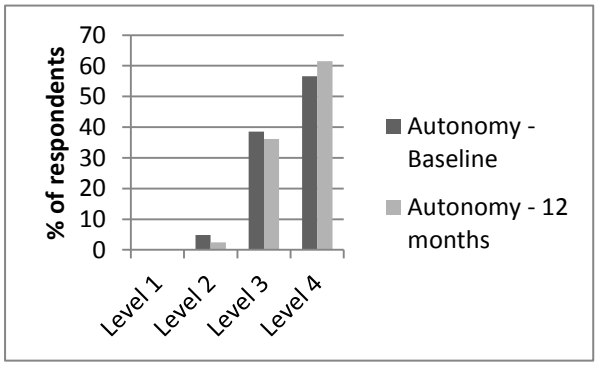
Stability Item



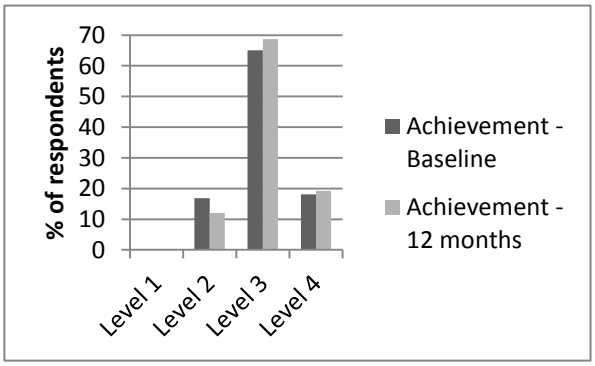
Attachment Item



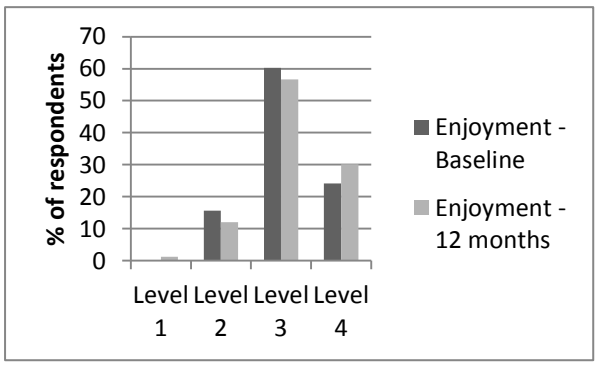
Autonomy Item



Achievement Item



Enjoyment Item



8.2.4.3. Non-weighted ICECAP-A score

Table 64 shows that cross-sectional correlations between the GAD-7 and the non-weighted ICECAP-A at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of the change in these measures between baseline and follow-up was weak and significant at the 1% level.

Table 64: Cross-sectional and change correlations between the GAD-7 and non-weighted ICECAP-A scores (n=335)

	ICECAP-A		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
GAD-7 score	-0.522**	-0.515**	-0.203**

* Significant at the 5% level, **Significant at the 1% level.

Table 65 shows change in non-weighted ICECAP-A scores by GAD-7 anchor change groups. In the group of participants reporting an improvement in GAD-7 scores, the mean non-weighted ICECAP-A score increased. In the group reporting a worsening of their GAD-7 scores, the mean non-weighted ICECAP-A score decreased. The change in ICECAP-A score was larger in the group that had worsened than in the group that had improved. Effect sizes for the group that improved were small, while for the group that worsened they were moderate.

The use of the non-weighted EQ-5D-3L as a reference measure (Table 66) shows differences from the non-weighted ICECAP-A analysis (Table 65). Change in the EQ-5D-3L scores as a percentage of possible change was larger for the group reporting an improvement in GAD-7 scores than for those reporting a worsening of scores. This is the reverse of what was found in the ICECAP-A analysis. The size of the change as a percentage of possible change, effect

sizes and SRMs, were larger for the ICECAP-A in the worsened group and larger for the EQ-5D-3L in the improved group..

Table 65: Mean change in non-weighted ICECAP-A scores by GAD-7 anchor change groups (n=335)

Anchor group	Baseline ICECAP-A scores	Follow-up ICECAP-A scores	Mean ICECAP-A change (95% CI)	Change as % of possible change	ES	SRM
Improved	16.012	16.390	0.378 (-0.024, 0.780)	2.5%	0.15	0.21
No change	17.430	17.569	0.139 (-0.05, 0.328)	0.9%	0.07	0.10
Worsened	16.442	15.279	-1.163** (-1.789, 0.537)	7.7%	0.53	0.57

* Significant at the 5% level, **Significant at the 1% level.

Table 66: Mean change in non-weighted EQ-5D-3L scores by GAD-7 anchor change groups (n=335) (for comparison)

Anchor group	Baseline EQ-5D-3L scores	Follow-up EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	8.16	7.308	-0.852** (-1.142, -0.561)	8.5%	0.56	0.65
No change	7.421	7.015	-0.406** (-0.56, -0.253)	4.1%	0.35	0.37
Worsened	7.878	8.097	0.219 (-0.69, 0.251)	2.2%	0.14	0.14

* Significant at the 5% level, **Significant at the 1% level.

8.2.4.4. ICECAP-A tariff score

Table 67 shows the cross sectional and change correlations between the GAD-7 and the ICECAP-A tariff. Moderate correlations were found at baseline and follow-up, which were statistically significant at the 1% level. The correlation of GAD-7 and ICECAP-A change scores between baseline and follow-up was weak and significant at the 1 % level.

Table 67: cross-sectional and change correlations between the GAD-7 and the ICECAP-A tariff (n=335)

	ICECAP-A		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
GAD-7 score	-0.515**	-0.509**	-0.205**

* Significant at the 5% level, **Significant at the 1% level.

Table 68 and Figure 27 show change in ICECAP-A tariff score by GAD-7 anchor change groups. In the group reporting an improvement in their GAD-7 scores, a small mean increase in ICECAP-A tariff scores was found. This change was not significant and the effect size and SRM were small. In the group of participants that reported an improvement in the GAD-7 scores, the mean ICECAP-A tariff score increased. This change was moderate and significant at the 1% level. Effect size and SRM were moderate. Comparison with the non-weighted ICECAP-O analysis shows that change as a percentage of possible change, the SRMs and the effects sizes are similar.

Table 68: Mean change in ICECAP-A tariff scores by GAD-7 anchor change groups (n=335)

Anchor group	Baseline ICECAP-A scores	Follow-up ICECAP-A scores	Mean ICECAP-A change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.844	0.864	0.02 (0.002 0.042)	2%	0.14	0.2
No change	0.913	0.917	0.004 (-0.003, -0.011)	0.4%	0.04	0.07
Worsened	0.863	0.792	-0.071** (-0.11, -0.032)	7.1%	0.58	0.55

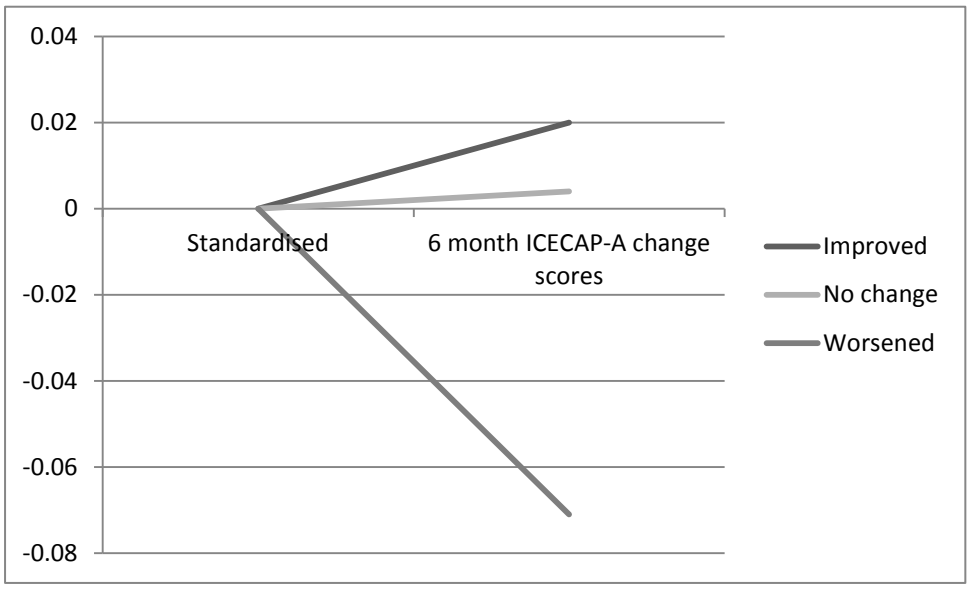
* Significant at the 5% level, **Significant at the 1% level.

Table 69: Mean change in EQ-5D-3L index scores by GAD-7 anchor change groups (n=335) (for comparison)

Anchor group	Baseline EQ-5D-3L scores	Follow-up EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.585	0.686	0.101** (0.052, 0.149)	6.3%	0.39	0.46
No change	0.667	0.717	0.05** (0.022, 0.078)	3.1%	0.24	0.24
Worsened	0.614	0.616	0.002 (-0.079, 0.083)	0.1%	0.01	0.01

* Significant at the 5% level, **Significant at the 1% level.

Figure 27: Mean change in ICECAP-A tariff scores by GAD-7 anchor change groups (n=355)



8.2.5. PHQ-8 anchor analysis summary

The full PHQ-8 anchor analysis is placed in Appendix 27, a summary is provided here.

In the groups reporting a worsening of PHQ-8 scores, a large reduction of 23 points was found in the percentage of respondents reporting full capability (level 4) for the Enjoyment item. Smaller changes of less than 12 percent were found for the other items. In the group reporting an improvement of PHQ-8 scores, smaller changes of less than 14 points were found in the percentage of respondents reporting full capability for each item.

Mean change in the ICECAP-A non-weighted score and the ICECAP-A tariff score was larger in the group reporting a worsening of PHQ-8 scores compared to those who reported an improvement. In the group that reported a worsening, the effect sizes and SRMs for the ICECAP-A change were small. The EQ-5D-3L (non-weighted and weighted scores) showed a different pattern of change to the ICECAP-A with larger changes in the group that reported improved PHQ-8 scores, compared to the group that reported worsened scores.

8.3. PastBP trial

8.3.1. Participant characteristics

The characteristics of the PastBP trial participants included in this responsiveness analysis are presented in Table 70. The participants from the PastBP trial were a majority male sample, with an average age of 71 years. The mean ICECAP-O scores of participants were slightly higher than values found in the general population [177]. Mean participant scores on the EQ-5D-3L and EQ-5D-3L VAS scores, of 0.75 and 74 respectively, were roughly equal to the UK national average for this age group [311]. On average, participants reported roughly 6 out of the possible 24 symptoms and side-effects assessed at both baseline and follow-up. The sub-scales of the SF-36 show that limitations due to physical function and physical role were the main impairments in this population. The cognitive functioning of the sample was good, with an MMSE score of 28, which indicates no substantial cognitive functioning problems within the sample [313]

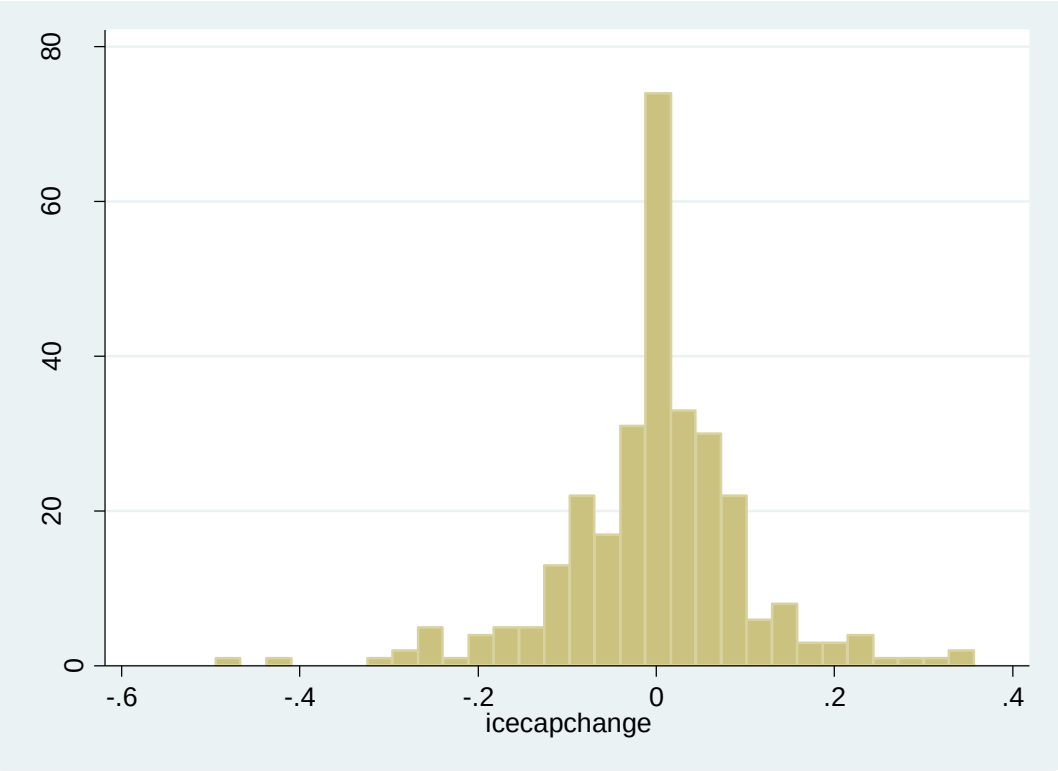
Table 70 shows that the mean score of the ICECAP-O did not change between baseline and follow-up. Figure 28 shows the range of individual change in ICECAP-O scores in the sample. The majority of the PastBP sample reported minimal change in their ICECAP-O scores of less than 0.1. This responsiveness analysis is therefore being completed with a population that reported a low level of change in their ICECAP-O tariff scores and have capability levels that are comparable to the general population.

Table 70: PastBP trial partcipant characteristics, including mean scores and median scores

Measures	Mean baseline values (SD)	Median baseline value (IQR)	Mean follow-up value (SD)	Median follow-up value (IQR)	Measure range	Sample size
Socio-demographic						
Age	71.4 (8.68)					303
Gender (% male)	62.05%					303
Health and functioning						
ICECAP-O tariff	0.86 (0.12)	0.89 (0.81,0.93)	0.86 (0.12)	0.89 (0.82,0.93)	0 to 1	301
EQ-5D-3L index	0.75 (0.23)	0.73 (0.69, 0.85)	0.76 (0.23)	0.76 (0.69,1)	-0.59 to 1	293
EQ-5D-3L VAS	74.3 (17.6)	78 (60,90)	75.4 (17.6)	80 (69,89)	0 to 100	286
MMSE	28.0 (2.22)	29 (27,30)	27.8 (2.54)	29 (27,30)	0 to 30	249
No. SSE (range 1 to 24)	5.95 (4.12)	6 (3,9)	5.69 (4.23)	5 (3,8)	0 to 24	164
SF-36						
• Physical function	42.4 (11.5)	45.4 (33.9,51.8)	40.7 (12.7)	44.4 (31.8,52.8)	14.9 to 57.0	251
• Physical role	41.1 (17.4)	47.0 (27.5,56.8)	30.9 (7.5)	27.5 (27.5,27.5)	17.7 to 56.8	91*
• Emotional role	45.6 (17.1)	55.9 (40.3,55.9)	45.6 (17.5)	55.9 (40.3,55.9)	9.2 to 55.9	276
• Vitality	49.7 (9.9)	50.84 (43.3,58.3)	49.8 (10.6)	50.8 (43.3, 58.3)	20.9 to 70.8	260
• Mental health	51.6 (9.2)	52.8 (46.1, 59.6)	52.6 (8.9)	53.9 (48.3,59.6)	14.5 to 64.1	264
• Social function	48.5 (10.40	51.4 (40.5,56.8)	48.3 (10.1)	51.4 (40.5, 56.8)	13.2 to 56.8	284
• Bodily Pain	48.6 (10.8)	48.0 (38.6,57.4)	48.9 (11.7)	52.7 (38.6,57.4)	19.8 to 62.1	266
• General health	45.3 (9.1)	47.2 (40.1,52.0)	46.4 (9.2)	47.2 (40.1,52.0)	16.2 to 63.9	268

* The sample size for the physical role sub-scale of the SF-36 is low. Manual inspection of the paper copies follow-up questionnaire packs indicates that this is likely due to participants missing an item of the scale because the item was placed after an "end of section" sub-heading, which encouraged participants to move onto the next section.

Figure 28: Frequency distribution of change in ICECAP-O tariff scores



8.3.2. Choice of anchors from PastBP trial

Table 71 shows the correlation of change scores (i.e. change in the scores between baseline and follow-up) between measures administered at baseline and follow-up in the PastBP trial. Change score correlations between most of the measures were weak, with only a small number of correlations over 0.3. The majority of correlations with the ICECAP-O change scores were under 0.2 and no correlation exceeded 0.3. The EQ-5D-3L VAS, the SF-36 general health sub-scale, the number of symptoms and side-effects, and the EQ-5D-3L index showed the strongest correlations with the ICECAP-O. With the exception of the EQ-5D-3L VAS and the SF-36 pain sub-scale, the strength of correlations between the EQ-5D-3L index score and other measures was comparable to the strength of correlations between the ICECAP-O tariff score and other measures.

Correlations of change scores between the ICECAP-O measure and other measures can be used to inform the selection of anchors for a responsiveness analysis. Based on these correlations and the considerations discussed in the methods chapter, the following anchors were selected for assessment of responsiveness in this chapter:

- An assessment using the EQ-5D-3L.
- An assessment using the EQ-5D-3L VAS (placed in appendices)
- An assessment using the SF-36 subscales of general health, vitality and social function (general health and vitality placed in appendices)
- An assessment using the Modified Rankin Scale.
- An assessment using the number of symptoms and side-effects that an individual is suffering from.

Table 71: Correlations between change scores of measures in the PastBP trial

	ICECAP -O tariff	EQ-5D-3L index	EQ-5D-3L VAS	Number of SSE	Phys func	Phys role	Emo role	Vitality	Emotion	Soc func	Pain	Gen health	MMSE
ICECAP-O tariff	1.00												
EQ-5D-3L Index	0.198	1.00											
EQ-5D-3L VAS	0.277	0.193	1.00										
Number of SSE	-0.235	-0.296	-0.269	1.00									
Physfunc	0.144	0.202	0.218	-0.203	1.00								
Phys role	0.071	0.124	0.134	-0.191	0.215	1.00							
Emo role	0.184	0.118	0.163	-0.048	0.159	0.130	1.00						
Vitality	0.19	0.198	0.353	-0.258	0.147	0.240	0.287	1.00					
Emotion	0.082	0.107	0.138	-0.052	0.130	0.055	0.233	0.286	1.00				
Social func	0.209	0.188	0.266	-0.352	0.258	0.245	0.273	0.357	0.177	1.00			
Pain	0.159	0.348	0.168	-0.362	0.234	0.290	0.162	0.339	0.093	0.510	1.00		
Gen Helath	0.254	0.201	0.284	-0.221	0.341	0.142	0.115	0.314	0.156	0.269	0.242	1.00	
MMSE	-0.119	-0.051	-0.212	0.142	-0.147	0.142	-0.082	-0.034	0.028	-0.139	-0.118	0.090	1.00

The Modified Rankin Scale is not included in this analysis as it is a 5 point categorical variable and therefore not appropriate for correlation analysis.

8.3.3. EQ-5D-3L index anchor analysis

8.3.3.1. Anchor group formation

Anchor groups were formed based on the same minimally important difference value of 0.074 taken from Walters and Brazier [262] as used in the BEEP analysis. Groups comprised of participants reporting equal to or greater than the minimally important change in EQ-5D-3L scores. Mean change in the EQ-5D-3L index as a percentage of possible change was roughly 15% in both change groups.

Table 72: Group numbers and mean EQ-5D-3L index change scores in the EQ-5D-3L index anchor groups (n=279)

Anchor group	Number in group	Mean EQ-5D-3L change in group (95% CI)	Mean EQ-5D-3L change as a % of possible change.
Improved	88	0.243 (0.210,0.276)	15.2%
No change	133	-0.004 (-0.009,0.001)	0.2%
Worsened	58	-0.245 (-0.288,-0.202)	15.4%

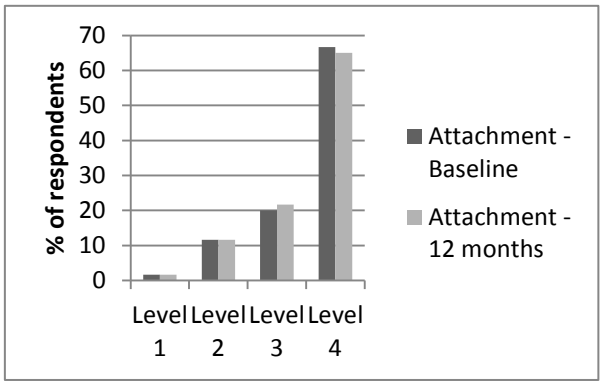
8.3.3.2. Item-by-item analysis

Figure 29 shows the response profiles at baseline and follow-up in the group of respondents who reported their EQ-5D-3L index scores worsening. These results are provided in numerical form in Appendix 30. The response profile of these respondents changed. Between baseline and follow-up there was a 22 point reduction in the percentage of respondents answering level 4 (full capability) for Control and a 9 or 10 point reduction in the

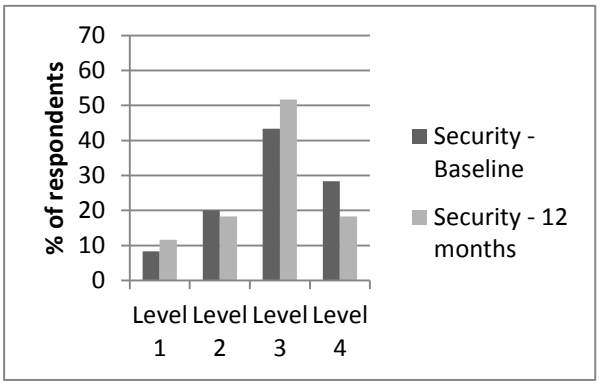
percentage in respondents answering level 4 for Security, Role and Enjoyment. In the Control, Role and Security items this change resulted in an increase or respondents answering level 3, while for Enjoyment an increase was seen in respondents answering level 2. The response profile for the Attachment item remained largely unchanged.

Figure 29: ICECAP-O response profile at baseline and follow-up for participants reporting a worsening of their EQ-5D-3L index scores

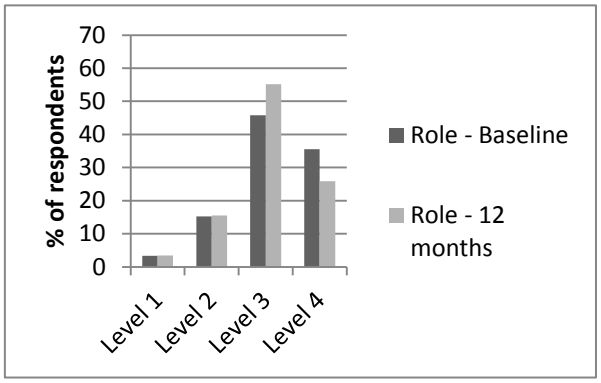
Attachment Item



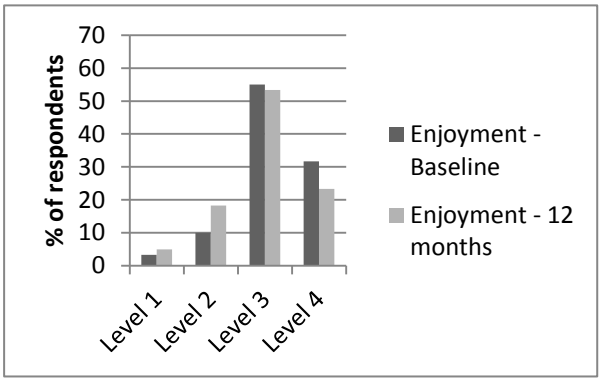
Security Item



Role Item



Enjoyment Item



Control Item

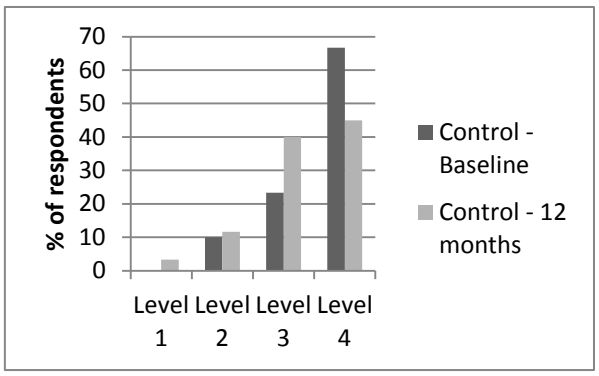
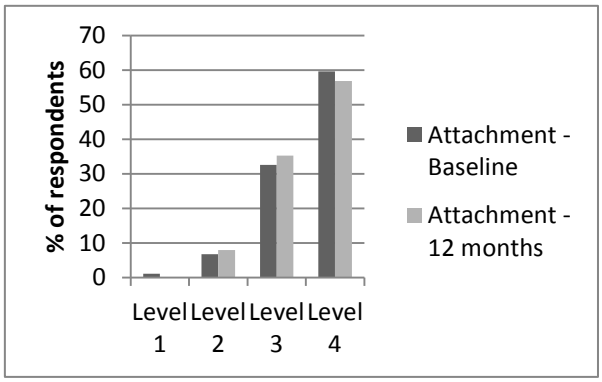


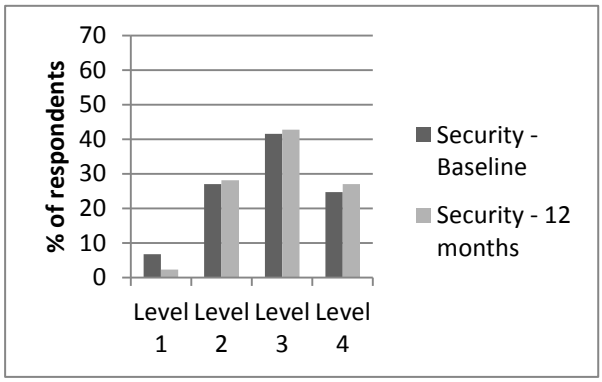
Figure 30 shows the ICECAP-O response profile at baseline and follow-up in the group of respondents reporting their EQ-5D-3L index scores improving. These results are provided in numerical form in Appendix 31. In comparison to the change in profile associated with a worsening of EQ-5D-3L scores, there were smaller, less pronounced changes in the ICECAP-O response profile of this group. There was a 10 point and 7 point increase between baseline and follow-up in the percentage of respondents answering level 4 (full capability) for Role and Enjoyment. There were small increases in the number of participants reporting the top level of Security and Control and a small decrease in the percentage of respondents reporting the top level of Attachment.

Figure 30: ICECAP-O response profile at baseline and follow-up for participants reporting an improvement in their EQ-5D-3L index scores

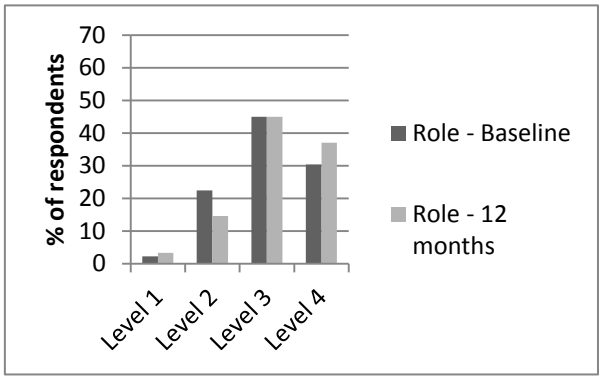
Attachment Item



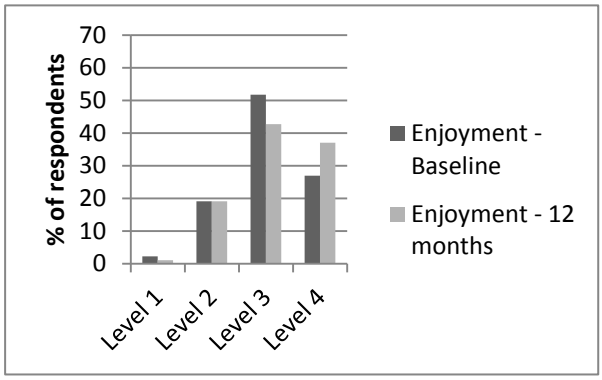
Security Item



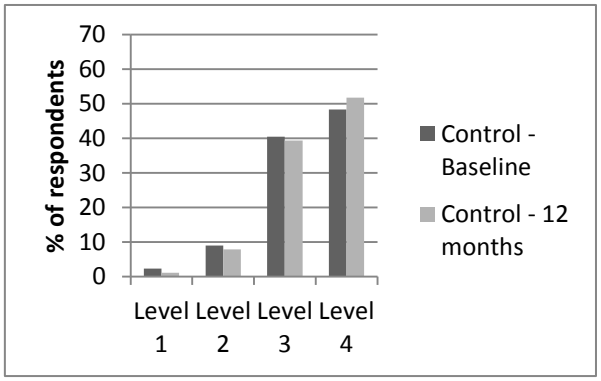
Role Item



Enjoyment Item



Control Item



8.3.3.3. Non-weighted ICECAP score analysis

Table 73 shows that cross-sectional correlations between the EQ-5D-3L index and the non-weighted ICECAP-O scores at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of the change in scores between baseline and 12-month follow-up in these measures was weak and significant at the 1% level.

Table 73: Cross-sectional and change correlations between the EQ-5D-3L index and non-weighted ICECAP-O scores (n=279)

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
EQ-5D-3L index score	0.548**	0.513**	0.205**

* Significant at the 5% level, **Significant at the 1% level.

Table 74 shows change in non-weighted ICECAP-O scores by EQ-5D-3L index anchor change groups. In the group of participants reporting an improvement in EQ-5D-3L index score, the mean non-weighted ICECAP-O score increased. In the group reporting a worsening of their EQ-5D-3L index score, the mean non-weighted ICECAP-O score decreased. The change in ICECAP-O score in the group that worsened was larger than in the group that improved. Table 72 shows the mean EQ-5D-3L change in these groups to be similar. Therefore, the proportion of change in ICECAP-O scores to EQ-5D-3L index change was larger in the group that worsened than in the group that improved. The effect size and SRM for ICECAP-O score changes were small.

Table 74: Mean change in non-weighted ICECAP-O scores by EQ-5D-3L anchor change groups (n=279)

Anchor group	Baseline ICECAP-O scores	Follow-up ICECAP-O scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	15.761	16.193	0.432 (-0.028, 0.891)	2.9%	0.15	0.2
No change	16.355	16.563	0.207 (-0.111, 0.526)	1.4%	0.08	0.11
Worsened	16.237	15.525	-0.712* (-1.284, -0.139)	4.8%	0.24	0.32

* Significant at the 5% level, **Significant at the 1% level.

8.3.3.4. ICECAP tariff analysis

Table 75 shows that cross-sectional correlations between the EQ-5D-3L index and the ICECAP-O tariff at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of change in these measures between baseline and 12 month follow-up was weak (approaching moderate) and statistically significant at the 1% level.

Table 75: Cross-sectional and change correlations between EQ-5D-3L index and ICECAP-O tariff (n=279)

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
EQ-5D-3L index score	0.458**	0.496**	0.179**

* Significant at the 5% level, **Significant at the 1% level.

Table 76 and Figure 31 show changes in ICECAP-O tariff score by EQ-5D-3L index anchor groups. In the group of participants reporting an improvement in EQ-5D-3L index scores the mean ICECAP-O tariff score increased. In the group reporting a worsening of their EQ-5D-

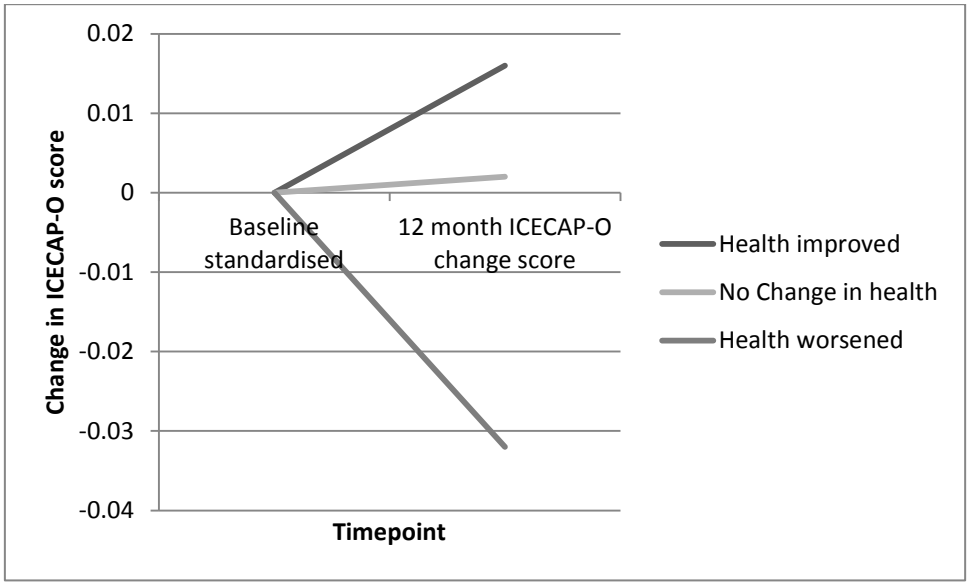
3L index score, the mean ICECAP-O tariff score decreased. The increase in ICECAP tariff scores in the group that improved was small, while the decrease in the group that worsened was moderate and statistically significant at the 5% level. As a proportion of possible change, change in the ICECAP-O tariff scores (Table 76), in both the improved and worsened groups, was smaller than in non-weighted ICECAP-O scores (Table 74). The effect sizes and SRMs were small to very small.

Table 76: Mean change in ICECAP-O tariff scores by EQ-5D-3L anchor change groups (n=279)

Anchor group	Baseline ICECAP-O scores	Follow-up ICECAP-O scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.850	0.866	0.016 (-0.007,0.039)	1.6%	0.13	0.15
No change	0.875	0.877	0.002 (-0.013,0.017)	0.2%	0.02	0.02
Worsened	0.862	0.830	-0.032* (-0.063,-0.001)	3.2%	0.22	0.27

* Significant at the 5% level, **Significant at the 1% level.

Figure 31: Mean change in ICECAP-O tariff scores by EQ-5D-3L anchor change groups (n=279)



8.3.4. EQ-5D-3L VAS anchor analysis summary

The full EQ-5D-3L VAS anchor analysis is placed in Appendix 32, a summary is provided here.

In the group that reported a worsening of EQ-5D-3L VAS scores, reductions of between 5 and 2 points were seen in the percentage of respondents reporting full capability on each item. In the group that reported an improvement, larger increases of between 4 and 10 percentage points were found in those reporting full capability for each item.

Mean change in the non-weighted ICECAP-O score and ICECAP-O tariff score was greater for the group that reported an improvement in their EQ-5D-3L VAS scores. Non-weighted ICECAP-O change as a percentage of possible change, effect sizes and SRMs were similar to those for the non-weighted EQ-5D-3L reference measure score. Change as a percentage of possible change was smaller for the ICECAP-O tariff than for the non-weighted ICECAP-O score.

8.3.5. Modified Rankin Scale anchor analysis

8.3.5.1. Anchor group formation

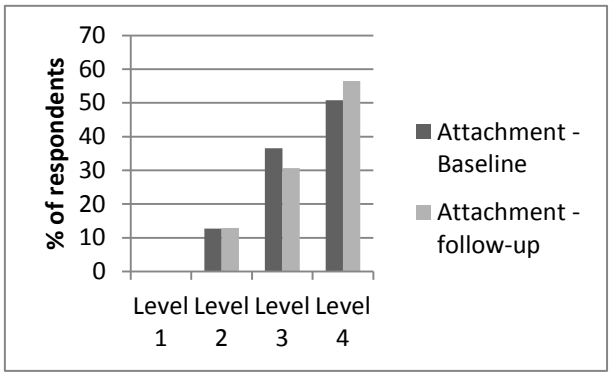
The anchor groups for the Modified Rankin Scale were formed using "naturally" occurring change groups of those who had improved or worsened by 1 or 2 points on the 5 point scale (no informant changed by more than 2 points). A change of -1 or -2 indicates an improvement in a participant's disability, while a change of +1 or +2 indicates a worsening of disability.

8.3.5.2. Item-by-item analysis

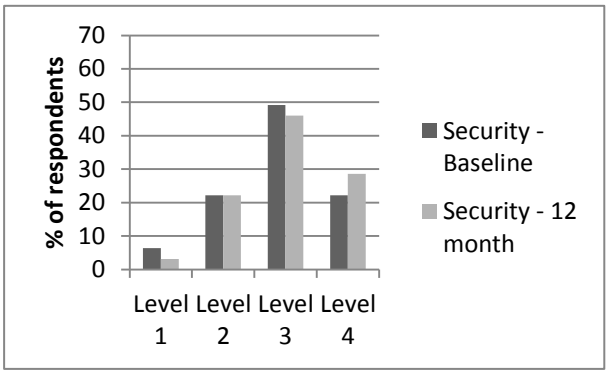
Figure 32 shows the response profiles at baseline and follow-up for respondents who reported an improvement in their Modified Ranking Scale scores (change of -1 or -2). These results are presented in a numerical form in Appendix 35. Between baseline and follow-up an increase of 15 percentage points was seen in participants answering level 4 (full capability) on Enjoyment. Changes in level 4 response of 1 and 7 percentage points were seen for the other items. There was also a notable reduction in the percentage of participants selecting the bottom two levels for Role, which resulted in an increase in those selecting the top two levels of this attribute.

Figure 32: ICECAP-O response profile at baseline and follow-up for participants reporting an improvement in their Modified Rankin Scale scores

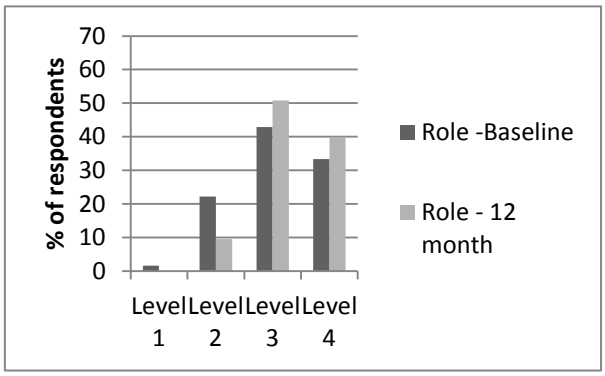
Attachment Item



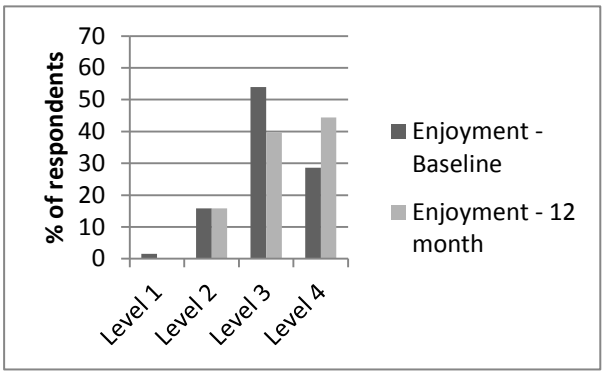
Security Item



Role Item



Enjoyment Item



Control Item

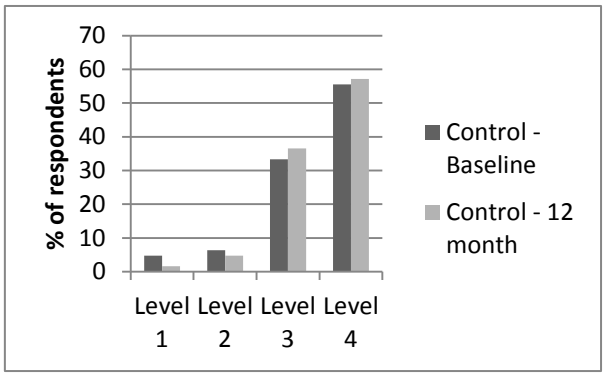
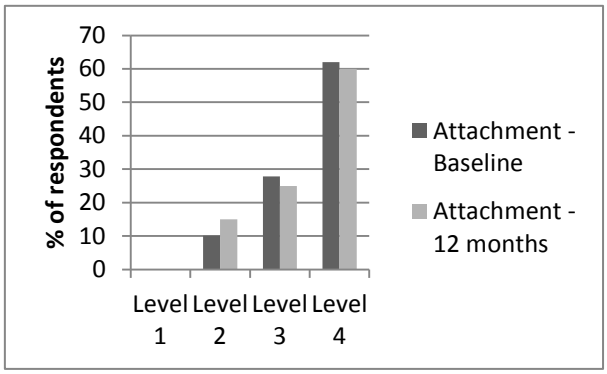


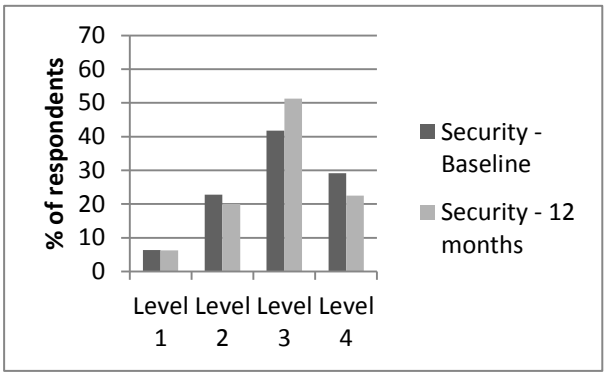
Figure 33 shows ICECAP-O response profiles at baseline and follow-up for those reporting a worsening of their Modified Rankin Scale scores (change of +1 or +2). These results are presented in numerical form in Appendix 36. Reductions in the percentage of respondents answering level 4 of 7 points for Security and Role and 9 points for Control were found. Smaller reductions of 2 and 5 points are seen in the percentage of respondents answering level 4 (full capability) of Attachment and Enjoyment. There is a notable increase of 9 points in the percentage of participants answering either of the bottom two levels for the Role item.

Figure 33: ICECAP-O response profile at baseline and follow-up for participants reporting a worsening of their Modified Rankin Scores

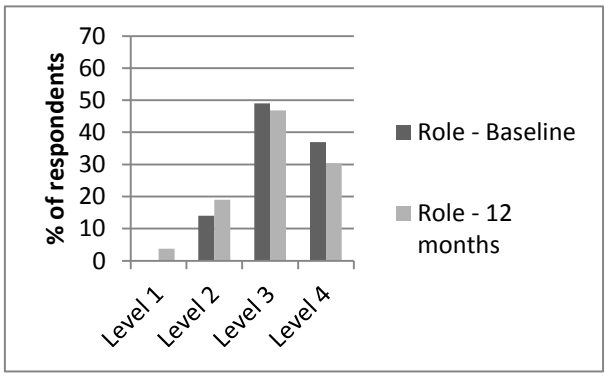
Attachment Item



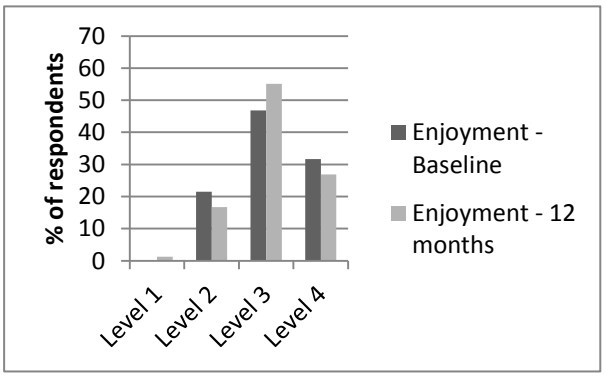
Security Item



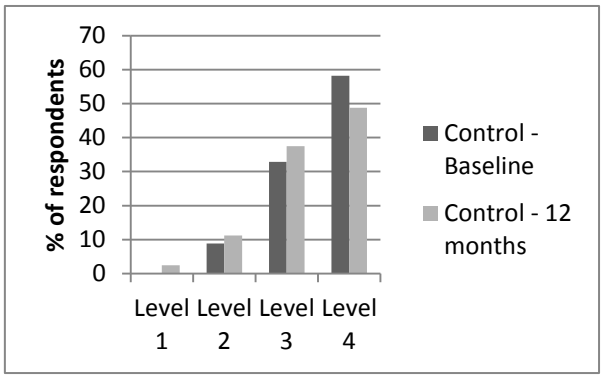
Role Item



Enjoyment Item



Control Item



8.3.5.3. Non-weighted ICECAP scores analysis

Table 77 shows change in non-weighted ICECAP-O scores by Modified Rankin Scale anchor groups. In the groups of participants reporting an improvement in their Modified Rankin Scale score (decrease in score) the mean non-weighted ICECAP-O scores increased. The increase in ICECAP scores in the group which had changed by 2 points on the Modified Rankin Scale was larger than in the group which had changed by 1 point. In the group of participants that reported a worsening of their Modified Rankin Scale score (increase in score) the mean non-weighted ICECAP-O score decreased. The decrease in the group which had changed by 1 point was smaller than in the group that had changed by 2 points. Changes in the groups that reported a worsening of their Modified Rankin Scale score were smaller than the changes in the group whose Modified Rankin Scale score had improved. Effect sizes and SRMs for the group reporting improvement in Modified Rankin Scale scores of 2 points (2 point decrease) were moderate, while in the group reporting a worsening by 2 points (2 point increase) they were large. Effect sizes and SRMs for the groups which changed by +/- 1 point on the Modified Rankin Scale were small to very small.

The use of the EQ-5D-3L non-weighted analysis (Table 78) as a reference measure shows differences to the non-weighted ICECAP-O score analysis (Table 77). For those reporting an improvement in Modified Ranking Scale scores of 2 points (2 point decrease), the effect size and SRM were larger for the non-weighted ICECAP-O scores than the non-weighted EQ-5D-3L scores. For those reporting a worsening of Modified Rankin Scale scores (2 point increase) the reverse was true. In the groups that reported an improvement or worsening by 1 point change as a percentage of possible change, effect sizes and SRMs were larger for the EQ-5D-3L non-weighted scores than for the ICECAP non-weighted scores. This indicates

differences between the measures in sensitivity to change in the Modified Rankin Scale by 1 point.

Table 77: Mean change in non-weighted ICECAP-O score by Modified Rankin Scale anchor change groups (n=288)

Anchor group	Number in group	Baseline ICECAP-O scores	Follow-up ICECAP-O scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
-2	7	14.571	17.142	2.571** (1.075, 4.068)	17.1%	1.53	1.58
-1	55	15.982	16.491	0.509* (0.074, 0.944)	3.3%	0.17	0.32
No change	149	16.127	16.161	0.033 (-0.798, 0.247)	0.2%	0.01	0.01
1	69	16.145	15.870	-0.275 (-0.798, 0.247)	1.8%	0.09	0.13
2	8	17.375	15.75	-1.625 (-3.670, 0.420)	10.8%	0.65	0.66

* Significant at the 5% level, **Significant at the 1% level.

Table 78: Mean change in non-weighted EQ-5D-3L score by Modified Rankin Scale anchor change groups (n=294) (for comparison)

Anchor group	Number in group	Baseline EQ-5D-3L scores	Follow-up EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
-2	8	8.875	7.125	-1.75* (-0.506,-2.994)	17.5%	0.81	1.18
-1	61	7.213	6.639	-0.574** (-0.201,-0.946)	5.7%	0.35	0.39
No change	143	7.049	6.839	-0.21 (0.001, -0.421)	2.1%	0.12	0.16
1	71	6.704	7.098	0.39** (0.113,0.676)	3.9%	0.25	0.33
2	11	5.545	6.909	1.364** (0.611,2.116)	13.6%	0.99	1.21

* Significant at the 5% level, **Significant at the 1% level.

8.3.5.4. ICECAP tariff analysis

Table 79 and Figure 34 shows change in ICECAP-O tariff score by Modified Rankin Scale anchor change groups. In the group of participants reporting a worsening of Modified Rankin Scale scores (increased score), the mean ICECAP-O tariff score decreased. In those reporting an improvement in Modified Rankin Scale scores (a decrease in scores) ICECAP-O tariff scores improved. For changes in both directions, ICECAP-O tariff score change was larger in the groups whose Modified Rankin Scale score had changed by 2 points than by 1 point. Effect sizes and SRMs were large for groups changing by 2 points and small for the 1 point change group.

In comparison to the non-weighted ICECAP-O score analysis (Table 77), change as a percentage of possible change is smaller in the ICECAP-O tariff analysis. This difference was particularly pronounced for the group who reported an improvement by 1 or 2 points on

the Modified Rankin Scale (a decrease of 1 or 2); less of a difference was found for those reporting a worsening. Using the EQ-5D-3L index score analysis as a reference measure (Table 80) shows that this trend is not present for the EQ-5D-3L index scores. Here, change as a percentage of possible change is largely unchanged between the non-weighted EQ-5D-3L scores and EQ-5D-3L index scores. Change as a percentage of possible change is larger in the EQ-5D-3L index analysis than in the ICECAP-O tariff. Non-systematic differences between the two measures in effect size and SRMs are seen, which may be a result of low numbers in the 2 point change groups.

Table 79: Mean change in ICECAP-O tariff score by change in Modified Rankin Scale score (n=288).

Anchor group	Number	Baseline ICECAP-O scores	Follow-up ICECAP-O scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
-2	7	0.803	0.891	0.088 ** (0.037,0.139)	8.8%	0.9	1.59
-1	55	0.852	0.877	0.025* (0.001,0.049)	2.5%	0.2	0.27
No change	149	0.863	0.864	0.001 (-0.016,0.018)	0.1%	0.01	0.01
1	67	0.867	0.847	-0.019 (-0.048,0.010)	1.9%	0.17	0.17
2	8	0.925	0.837	-0.088* (-0.169,-0.006)	8.8%	1.35	0.9

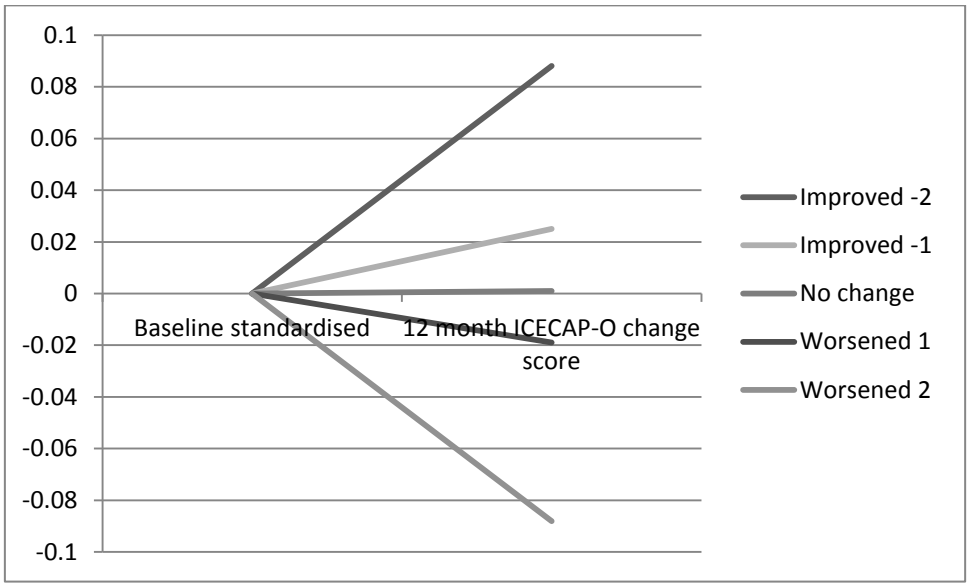
* Significant at the 5% level, **Significant at the 1% level.

Table 80: Mean change in EQ-5D-3L index score by change in Modified Rankin Scale score (n=294) (for comparison)

Anchor group	Number	Baseline scores	12 month follow-up scores	Mean change (95% CI)	Change as % of possible change	ES	SRM
-2	8	0.453	0.735	0.282* (0.022,0.542)	17.7%	0.76	0.91
-1	61	0.718	0.798	0.08** (0.029,0.130)	5%	0.37	0.41
No change	143	0.74	0.769	0.029 (-0.006, 0.065)	1.8%	0.13	0.13
1	71	0.776	0.742	-0.034 (-0.075,0.008)	2.1%	0.14	0.19
2	11	0.922	0.687	-0.235* (-0.415, -0.055)	14.7%	2.06	0.88

* Significant at the 5% level, **Significant at the 1% level.

Figure 34: Mean change in ICECAP-O tariff score by change in Modified Rankin Scale score (n=288)



8.3.6. SF-36 subscale anchor analysis

The three SF-36 subscales of general health, vitality and social functioning were selected for use as anchors in the responsiveness analysis. These anchors showed strong cross-sectional correlations with the ICECAP-O tariff at baseline and follow up and showed the strongest change score correlations between the ICECAP-O tariff and any of the SF-36 scales. The general health and vitality sub-scales analyses are placed in the appendices. A summary of results is included below. The anchor groups for the SF-36 scales were taken from the User's Manual for the SF-36v2 Health Survey [288].

8.3.7. SF-36 general health sub-scale analysis

The full analysis of the SF-36 general health sub-scale is provided in Appendix 37, a summary is provided here.

In the group of respondents that reported a worsening of their general health, sub-scale score reductions of between 17 and 8 points were seen in the percentage of respondents reporting full capability on each item, with the largest reductions being in Security and Enjoyment. In the group reporting an improvement in their general health sub-scale scores, increases of between 12 and 17 points were found in the percentage reporting full capability on each item, apart from control which showed minimal change.

Mean change in both the non-weighted ICECAP-O scores and the ICECAP-O tariff was larger in the group reporting an improvement in their general health sub-scale scores than those reporting a worsening of scores (which showed little change). Effects sizes and SRMs were small or moderate for the improved group. The changes in ICECAP-O scores in the

improvement group were larger than the changes in EQ-5D-3L scores, while the reverse was true for those reporting a worsening of general health.

8.3.8. SF-36 vitality sub-scale analysis

The full analysis of the vitality sub-scale is presented in Appendix 40, a summary is provided here.

In the group reporting a worsening of vitality, response profiles showed a 6 to 9 percentage point reduction in respondents answering level 4 (full capability) on each item. In those reporting an improvement in vitality, an increase of greater than 10 percentage points was seen for people answering the top level of Attachment, Role and Enjoyment items.

Mean changes in both the non-weighted ICECAP-O scores and the ICECAP-O tariff were larger in the group of respondents who reported improved vitality in comparison to those reporting worsened vitality. In the improved group, effect sizes were small and SRMs were moderate. The non-weighted and index score EQ-5D-3L analyses showed smaller change as a percentage of possible change, effect sizes and SRMs than the ICECAP-O for the group of respondents whose vitality had improved. Change was larger in the EQ-5D-3L analyses than the ICECAP-O analyses for the group of respondents whose vitality had worsened. ICECAP-O change as a percentage of possible change was smaller in the tariff analysis than the non-weighted analysis.

8.3.9. SF-36 social function sub-scale analysis

8.3.9.1. Anchor group formation

Anchor groups were formed of participants reporting a change equal to or greater than the minimally important change of 6.2 taken from the SF-36 User's Manual [288]. The mean change each groups was roughly 16% of the possible change on the measure.

Table 81: Numbers in group and the mean change in SF-36 social function sub-scale scores in anchor groups (n=267)

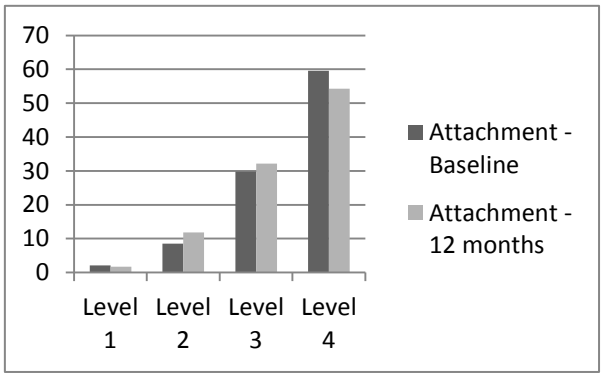
Anchor group	Number	Mean SF-36 social function change in group (95% CI)	Mean SF-36 social function sub-scale change as a % of possible change
Improved	40	16.48 (14.334, 18.625)	16.5%
No change	180	0.055 (-0.398, 0.508)	0.06%
Worsened	47	-15.622 (-17.362, -13.882)	15.6%

8.3.9.2. Item-by-item analysis

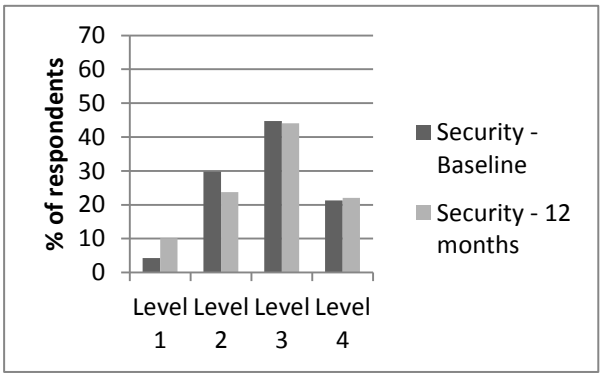
Figure 35 shows the response profiles at baseline and follow-up for respondent who reported a worsening of their EQ-5D-3L scores. These results are presented in numerical form in Appendix 43. The response profiles changed, with reductions of 5 to 14 points in the percentage of respondents reporting full-capability on each item. The exception was Security, which showed little change in the top two levels of the item, but showed notable change in the bottom two levels, with an increase of 6 points in the percentage of respondents reporting no capability in the measure.

Figure 35: ICECAP-O response profiled for participants reporting a worsening in their SF-36 social function sub-scale score.

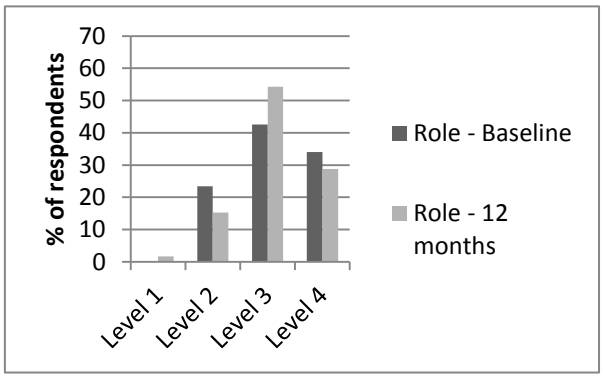
Attachment Item



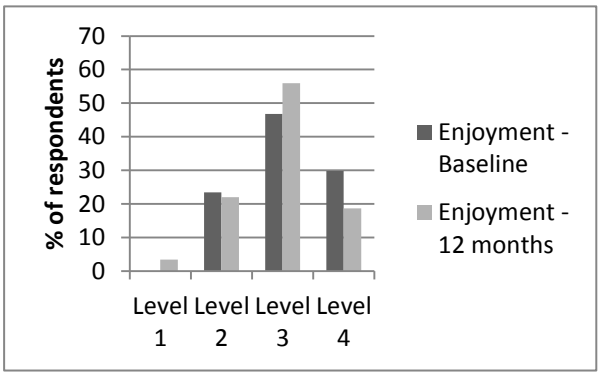
Security Item



Role Item



Enjoyment Item



Control Item

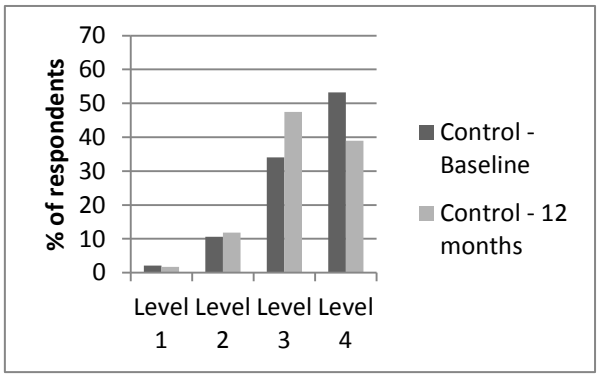
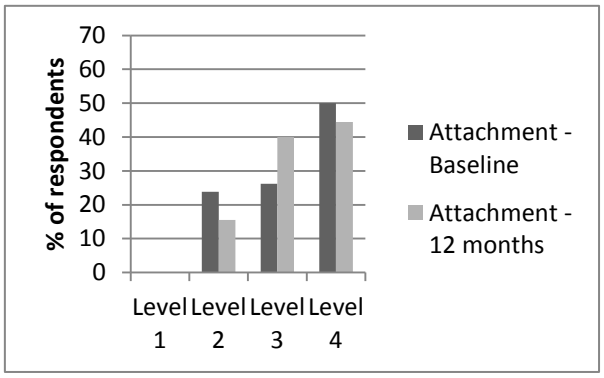


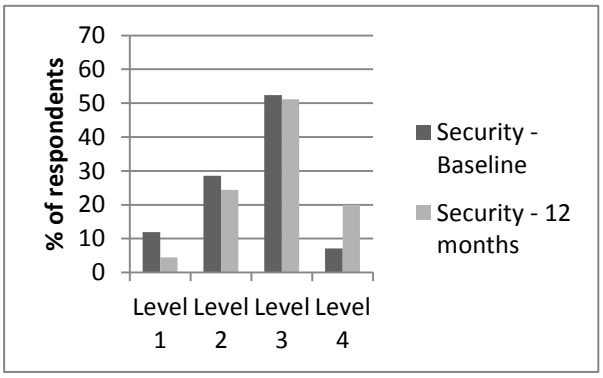
Figure 36 shows the response profile at baseline and follow-up for the group of respondents reporting an improvement in their SF-36 social function sub-scale scores. These results are reported in numerical form in Appendix 44. Increases of 10 and 17 points in the percentage of respondents reporting full capability were seen for Role and Enjoyment items, while a reduction of 6 percentage points was found for Attachment. There were also noticeable changes at lower levels of capability with reductions of 5 to 9 points in the percentage of respondents selecting level two on Attachment, Security and Role items.

Figure 36: ICECAP-O response profiles for participants reporting an improvement in their SF-36 social function sub-scale score.

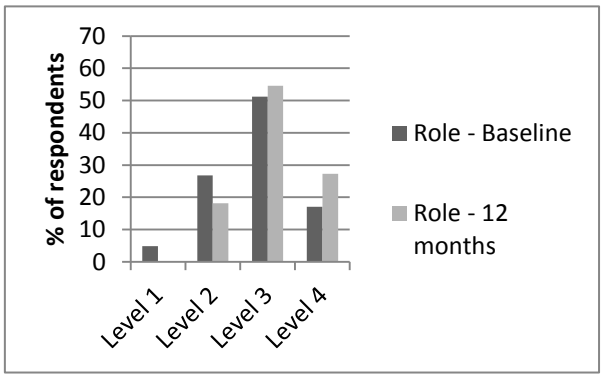
Attachment Item



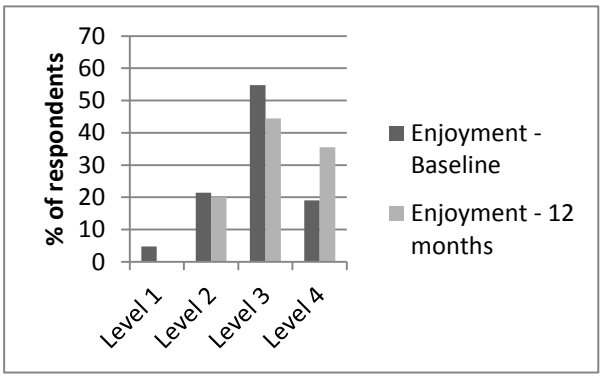
Security Item



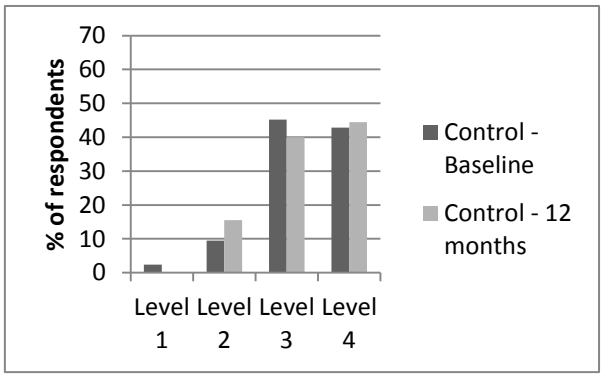
Role Item



Enjoyment Item



Control Item



8.3.9.3. Non-weighted ICECAP-O scores analysis

Table 82 shows that moderate cross-sectional correlations, statistically significant at the 1% level, between the ICECAP-A non-weighted scores and the SF-36 social function sub-scale scores, were found at baseline and follow-up. The correlation between change scores on both measures was weak and significant at the 1% level.

Table 82: Cross-sectional and changes correlations between SF-36 social function scale and non-weighted ICECAP-O score (n=267).

	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
SF-36 Social Function	0.49**	0.5**	0.21**

**Difference significant at p<0.01

Table 87 shows the changes in non-weighted ICECAP-O scores by SF-36 social functioning sub-scale anchor groups. In the group reporting an improvement in social function scores the mean non-weighted ICECAP-O score increased. In the group reporting a worsening in social function scores, the non-weighted ICECAP-O scores decreased. The increase in non-weighted ICECAP-O scores was larger than the decrease. Effect size and SRM for the group that improved were small to medium, while for the group that worsened they were very small. The non-weighted EQ-5D-3L analysis presented in Table 84 shows small differences in the ICECAP-O non-weighted analysis. Change as a percentage of possible change, effect size and SRM in the group that improved were slightly larger for the ICECAP-O than for the EQ-5D-3L. The reverse was observed in the group that reported a worsening in social function scores.

Table 83: Mean change in non-weighted ICECAP-O score by SF-36 social function scale change (n=267)

Anchor group	Baseline ICECAP-O scores	12 month ICECAP-O follow-up scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	14.725	15.85	1.125** (0.445, 1.805)	7.5%	0.41	0.53
No change	16.527	16.593	0.066 (-0.209, 0.341)	0.4%	0.02	0.03
Worsened	15.851	15.553	-0.298 (-0.945, 0.350)	1.9%	0.11	0.13

** Difference significant at p<0.01

Table 84: Mean change in non-weighted EQ-5D-3L scores by SF36 social-function health sub-scale change (n=281) (for comparison)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	7.932	7.25	-0.682** (-1.233,-0.13)	6.8%	0.35	0.37
No change	6.78	6.599	-0.187 (-0.361,-0.012)	1.9%	0.11	0.16
Worsened	7.127	7.382	0.255 (-0.128,0.638)	2.5%	0.14	0.18

** Difference significant at p<0.01

8.3.9.4. ICECAP-O tariff analysis

Table 85 shows cross-sectional correlations between the ICECAP-O tariff and SF-36 social function sub-scale at baseline and follow-up that were moderate and statistically significant at the 1% level. The cross-sectional correlation between these scores was weak and statistically significant at the 1% level.

Table 85: Cross-sectional and change correlations between SF-36 social function scale and the ICECAP-O tariff (n=267).

	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
SF-36 Social Function	0.49**	0.5**	0.21**

**Difference significant at p<0.01

Table 86 and Figure 37 show the changes in the ICECAP-O tariff score by SF-36 social function sub-scale anchor groups. In the group reporting an improvement in social function scores, the mean ICECAP-O tariff scores improved. This change was statistically significant at the 1% level. In the group that reported a worsening of social function scores, there was little change in the ICECAP-O tariff score. The effects size for the group that improved was small, while the SRM was moderate.

In comparison to the non-weighted ICECAP-O score, change as a percentage of possible change was smaller for the ICECAP-O tariff scores. There are greater differences between the ICECAP-O and EQ-5D-3L in the value weighted analysis than the non-weighted analysis in change as a percentage of possible, effect size and SRM.

Table 86: Mean change in ICECAP-O tariff score by SF-36 social function scale change (n=267)

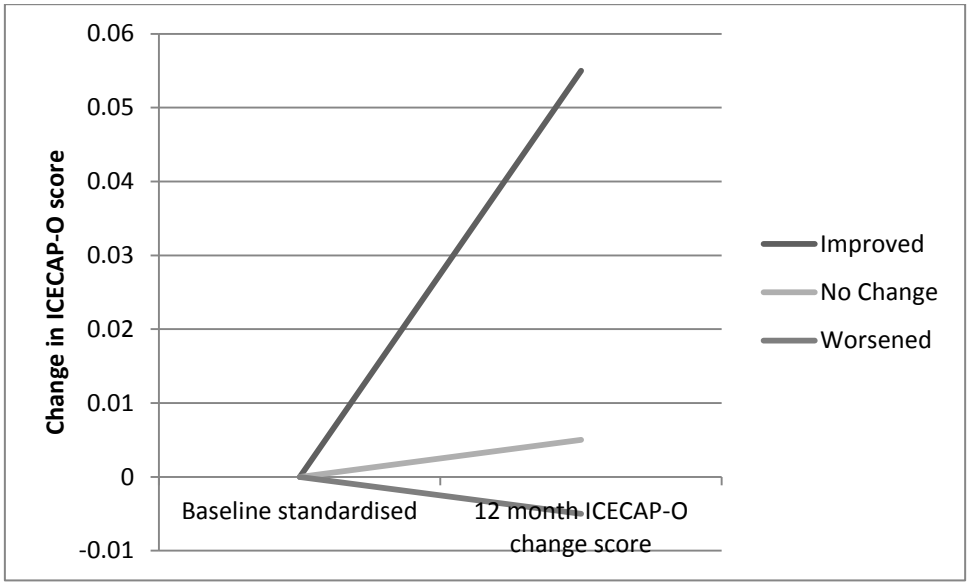
Anchor group	Baseline scores	12 month follow-up scores	Mean change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.799	0.854	0.055** (0.09, 0.02)	5.5%	0.39	0.55
No change	0.88	0.875	0.005 (0.009, 0.019)	0.5%	0.03	0.04
Worsened	0.849	0.844	-0.005 (-0.04, 0.029)	0.5%	0.04	0.04

** Difference significant at p<0.01

Table 87: Mean change in EQ-5D-3L index scores by SF36 social-function health sub-scale change (n=281) (for comparison)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.637	0.707	0.07 (-0.029,0.169)	4.4%	0.29	0.21
No change	0.769	0.801	0.032 (0.005, 0.059)	2%	0.14	0.18
Worsened	0.722	0.687	-0.035 (-0.097, 0.027)	2.2%	0.14	0.15

Figure 37: Mean change in ICECAP-O tariff score by SF-36 social function sub-scale change



8.3.10. Symptoms and side effects anchor analysis

8.3.10.1. Anchor group formation

No psychometric literature exists on the minimally important difference for the 24 item symptoms and side effects questionnaire used in the PastBP trial questionnaire pack. As some of the conditions assessed were relatively minor, such as sore eyes, while some were more serious, such as impotence, fatigue or nausea, it was thought inappropriate to assume that the increase or reduction in one symptom or side effect would represent a minimally important difference to participants. Therefore, groups were formed using the inter-quartile range values of +/-2 side effects. Table 88 includes the mean change in symptoms and side-effects: the improved group had a mean reduction in side effects of 3.6, while those in the worsened group had an increase in side-effects of 4.3. Change in number of symptoms and side-effects as a percentage of possible change, was similar in both groups.

Table 88: Numbers in groups and mean SSE change scores in SSE anchor change groups (n=107)

Anchor group	Number in group	Mean SSE change in group (95% CI)	SSE change as a % of possible change
Improved	35	-3.666 (-4.425, -2.906)	15.3%
No change	46	0.113 (-0.094, 0.320)	0.4%
Worsened	26	4.312 (3.038, 5.585)	15.4%

The symptoms and side-effects questionnaire is a multi-item questionnaire assessing a broad range of symptoms. To provide indications of the main "drivers" of change in the groups that

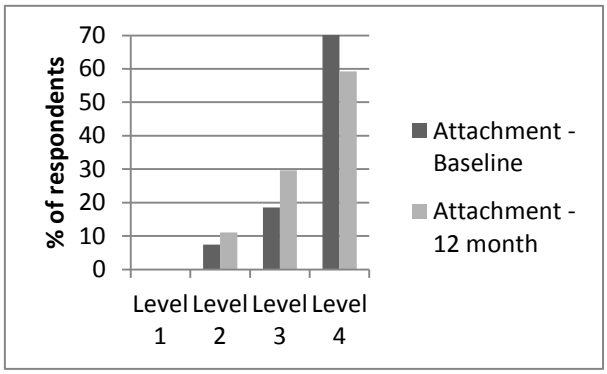
reported an increase or decrease in symptoms and side effects, the percentage change, between baseline and follow-up, in the number of participants reporting each individual side-effect is reported in Appendix 45. This shows that pain, breathlessness, fatigue and sleep difficulties show reductions of over 30% in respondents in the group reporting an overall improvement in their score of two or more. Pain, breathlessness, fatigue, loss of strength and leg and ankle swelling show increases approaching or in excess of 30% of respondents in the group reporting an overall worsening of scores of two or more.

8.3.10.2. Item-by-item analysis

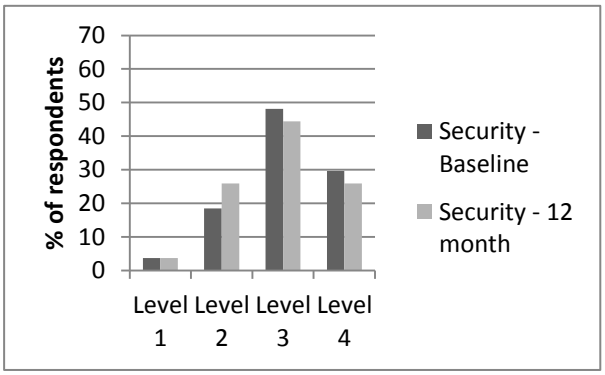
Figure 38 shows the response profiles at baseline and follow-up for the group of respondents who reported a worsening of their symptoms and side-effects (an increase by 2 or more). These results are presented in numerical form in Appendix 46. There are reductions of 15, 19 and 10 points in the percentage of participants answering level 4 (full capability) of Attachment, Role and Enjoyment, respectively. Smaller reductions of 4 were found points in the percentage of people reporting the top levels of Security and Control.

Figure 38: ICECAP-O response profile at baseline and follow-up for participants reporting a worsening in number of SSE

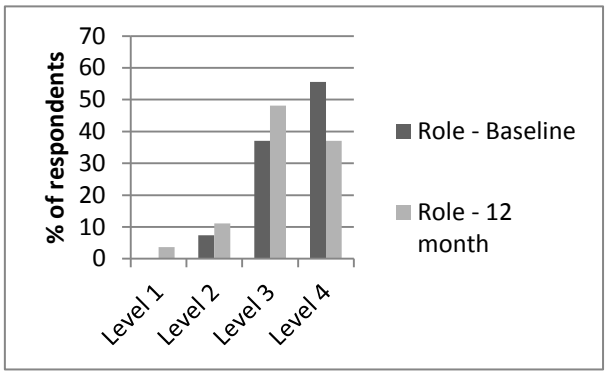
Attachment Item



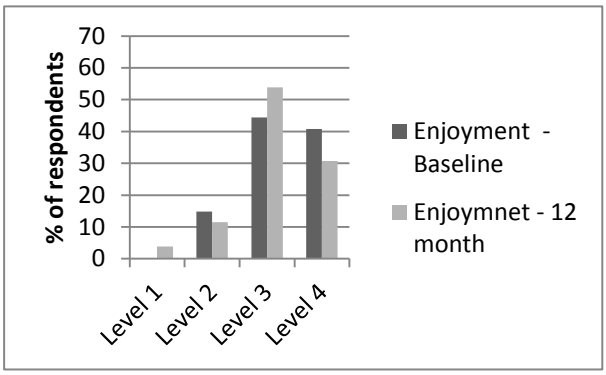
Security Item



Role Item



Enjoyment Item



Control Item

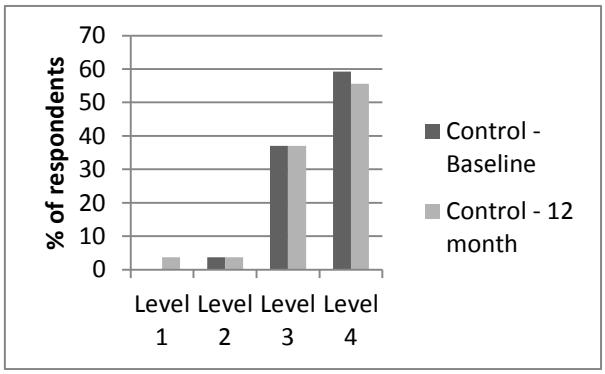
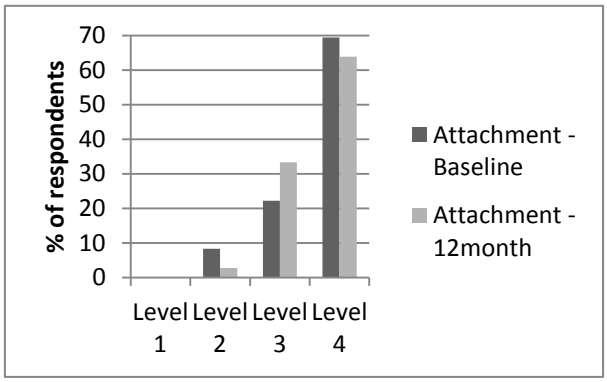


Figure 39 shows the response profile of participants who reported a reduction in symptoms and side-effects. These results are presented in numerical form in

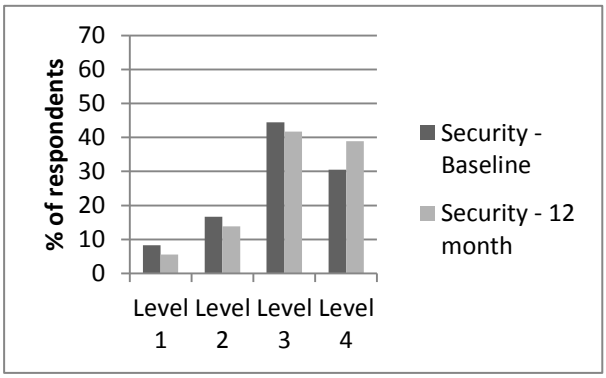
Appendix 47. Increases of 15 points and 14 points were found in the percentage of participants answering level 4 (full capability) for Role and Enjoyment, respectively. A smaller increase of 8 points was found in the percentage of participants reporting level 4 for Security. A 14 percentage point reduction in the number of participants reporting level 4 for Control was seen. A smaller reduction was found in the Attachment item. Therefore, the response profile for an improvement in symptoms and side-effects shows change in different directions for different items.

Figure 39: ICECAP-O response profile at baseline and follow-up for participants reporting an improvement in number of SSE

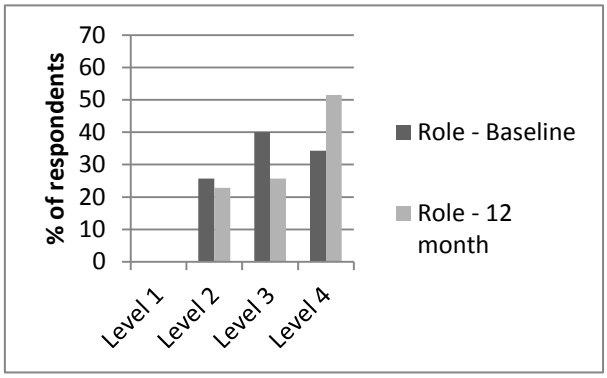
Attachment Item



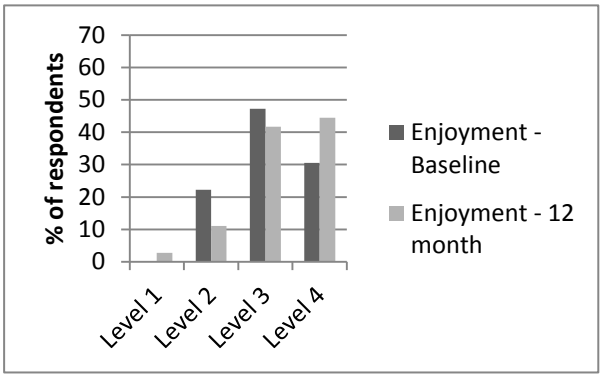
Security Item



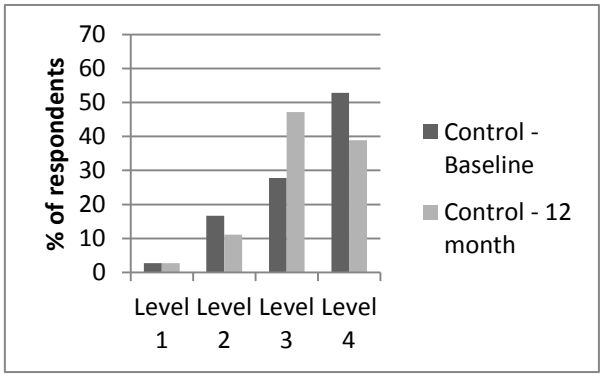
Role Item



Enjoyment Item



Control Item



8.3.10.3. Non-weighted ICECAP score analysis

Table 89 shows the correlations between the non-weighted ICECAP-O scores and symptoms and side-effects questionnaire at baseline, follow-up and over time. The cross-sectional correlations at both baseline and follow-up are moderate and significant at the 1% level. The correlation between the change scores of the measures is weak and significant at the 1% level.

Table 89: Cross-sectional and change correlations SSE and non-weighted ICECAP-O scores (n=107)

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
Symptoms and side-effects questionnaire	-0.497**	-0.439**	-0.299**

* Significant at the 5% level, **Significant at the 1% level.

Table 90 shows that the group of participants reporting an improvement in symptoms and side effects (a reduction of 2 or more), reported an increase in non-weighted ICECAP-O scores. The group of participants reporting a worsening of symptoms and side effects (an increase of 2 or more), reported a reduction in non-weighted ICECAP-O scores. The mean change for those reporting a worsening of symptoms and side-effects was larger than for those reporting an improvement. Effect sizes and SRM were small.

Differences between the non-weighted EQ-5D-3L reference analysis (Table 91) and non-weighted ICECAP-O analysis (Table 90) can be seen. Change as a percentage of possible change on the EQ-5D-3L was larger for the group of participants reporting an improvement of symptoms and side-effects, than for the group reporting a worsening. This is the opposite of the response pattern seen in ICECAP-O, where the change for the worsened group was larger

than for the improved. Effect sizes and SRMs for the EQ-5D-3L were larger than for the ICECAP-O in the group that improved and similar for the group that worsened..

Table 90: Mean change in non-weighted ICECAP-O scores by SSE anchor change groups (n=107)

Anchor group	Baseline ICECAP-O scores	12 month follow-up ICECAP-O scores	Mean change in ICECAP-O scores (95% CI)	Change as % of possible change	ES	SRM
Improved	15.971	16.514	0.543 (-0.071, 1.157)	3.6%	0.19	0.3
No change	16.26	16.217	-0.043 (-0.512, 0.425)	0.2%	0.01	0.02
Worsened	17	16.222	-0.777 (-1.797, 0.242)	5.2%	0.25	0.3

* Significant at the 5% level, **Significant at the 1% level.

Table 91: Mean change in non-weighted EQ-5D-3L scores by SSE anchor change groups (n=115) (for comparison)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	7.475	6.775	-0.7** (-1.13, -0.27)	7%	0.39	0.52
No change	6.614	6.386	-0.227 (-0.573,0.119)	2.3%	0.13	0.2
Worsened	7	7.322	0.322 (-0.216,0.861)	3.2%	0.25	0.22

* Significant at the 5% level, **Significant at the 1% level.

8.3.10.4. ICECAP tariff analysis

Table 92 shows that the baseline and 12 month follow-up cross-sectional correlations between the ICECAP-O tariff and the number of symptoms and side-effects were moderate and

statistically significant at the 1% level. The correlation in change scores of these measures was weak and statistically significant at the 5% level.

Table 92: Cross-sectional and change correlations between SSE and ICECAP-O measure (n=107)

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	Follow-up	
SSE	-0.425**	-0.385**	-0.235*

* Significant at the 5% level, **Significant at the 1% level.

Participants reporting a worsening of symptoms and side-effects had a reduction in ICECAP-O tariff scores (Table 93 and Figure 40). Participants reporting an improvement of symptoms and side-effects had an increase in ICECAP-O capability scores. These changes were small and neither were statistically significant at the 5% level. Effect sizes and SRMs were small. Changes as a percentage of possible change was smaller in the ICECAP-O tariff analysis than in the non-weighted ICECAP-O analysis. The difference in scores between the weighted and non-weighted analysis was also seen in the EQ-5D-3L.

Table 93: Mean change in ICECAP-O tariff score by SSE anchor change groups (n=107)

Anchor group	Baseline ICECAP-O scores	12 month follow-up ICECAP-O scores	Mean change in ICECAP-O scores (95% CI)	Change as % of possible change	ES	SRM
Improved	0.855	0.878	0.023 (-0.012, 0.057)	2.3%	0.16	0.22
No change	0.864	0.861	-0.003 (-0.030, 0.023)	0.3%	0.03	0.04
Worsened	0.897	0.858	-0.039 (-0.091, 0.014)	3.9%	0.44	0.3

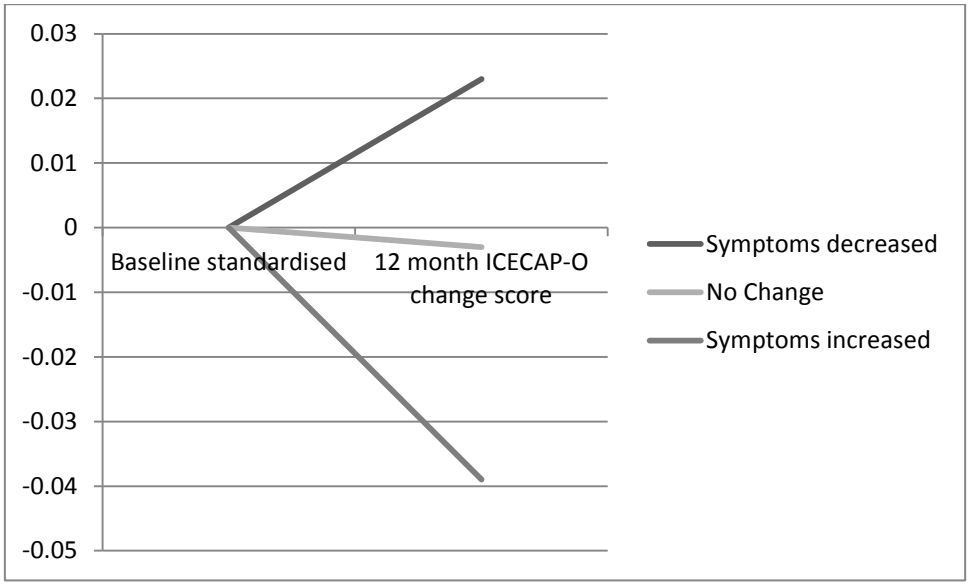
* Significant at the 5% level, **Significant at the 1% level.

Table 94: Mean change in EQ-5D-3L index scores by SSE anchor change groups (n=115) (for comparison)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.692	0.777	0.084** (0.024,0.145)	5.3%	0.38	0.45
No change	0.811	0.835	0.024 (-0.032, 0.079)	1.5%	0.11	0.13
Worsened	0.747	0.714	-0.033 (-0.12, 0.053)	2.1%	0.14	0.14

* Significant at the 5% level, **Significant at the 1% level.

Figure 40: Mean change in ICECAP-O tariff score by SSE anchor change groups (n=107)



8.4. Trends in the responsiveness analysis

The comparisons made in this chapter between the analyses using the non-weighted ICECAP scores and the ICECAP tariff scores, and between the ICECAP measures and the EQ-5D-3L have allowed the identification of a number of trends. These trends are summarised here. They are discussed in greater detail in Chapter 9 in the context of the ICECAP measure development and evaluation literature.

8.4.1. Small to moderate changes and effect sizes

Throughout the chapter, change in the ICECAP tariff, and the effect sizes and SRMs were small or moderate. Change in the ICECAP-A and ICECAP-O tariff scores frequently fell between 0.02 and 0.04. The notable exceptions to this were change in the ICECAP measures in response to change in the Modified Rankin Scale (ICECAP-O) and change on the psychological health measure of the GAD-7 and PHQ-8 (ICECAP-A). Small or moderate effect sizes and SRMs indicate that the ratio of change to standard deviation (either at baseline or overtime) was small: put another way, the signal to noise ratio was small.

As described in Chapter 3, the potential for ceiling effects could not be fully assessed due to the lack of a comparator capability measure. However, response profiles provide some indication that the small changes in ICECAP scores, when respondents report an improvement in health, may be due to a ceiling effect in some items. In both the ICECAP-O and the ICECAP-A over 50% of respondents frequently answer the top level (level 4) for the Attachment item and the Control or Autonomy item. In respondents whose health has improved, change profiles show small changes on these items between baseline and follow-

up. This is an indication that a ceiling effect may be reducing the sensitivity of the measure in those who reported a high level of capability at baseline.

8.4.2. Smaller changes in scores in tariff analyses

In the majority of anchor analyses presented in this chapter, the ICECAP tariff score showed smaller changes over time than the analyses using the non-weighted ICECAP score. This is apparent through considering change as a percentage of possible change on that measure, which standardises change to allow this comparison to be made. In the ICECAP-O analyses using EQ-5D-3L, the EQ-5D-3L VAS, the Modified Rankin Scale, the SF-36 General Health sub-scale, the Symptoms and Side-effects as anchors and the ICECAP-A analyses using the EQ-5D-3L and PHQ-8 as anchors, change as a percentage of possible change was smaller in the tariff analysis than in the non-weighted analysis. This indicates that, in a population with high initial levels of capability, when the general population values were applied to the "raw" non-weighted scores, the size of change on the ICECAP measures was suppressed.

8.4.3. Similar responsiveness of the EQ-5D-3L and ICECAP measures

The use of the EQ-5D-3L as a comparator on anchor measures, allowed the change as a percentage of possible change, effect sizes and SRMs of the ICECAP-O to be placed in context of scores on another value (or preference) weighted measure. When change was standardised as a percentage of possible change, it was apparent that change in the non-weighted analyses of the ICECAP measures was similar to change in the non-weighted analyses of the EQ-5D-3L in the analyses using the EQ-5D-3L VAS, Modified Rankin Scale, symptoms and side-effects, GAD-7 and PHQ-8. When the general population value weights

were applied to the ICECAP measures and the preference weights were applied to the EQ-5D-3L measure, the difference between the change scores of the two measures increased. This was due to the value weights supressing change to a greater extent on the ICECAP measures than the EQ-5D-3L. This in turn was likely to be because individuals were concentrated at the top end of capability, where relatively little additional value is obtained from a shift between ICECAP levels, compared with change at lower levels of capability (discussed in greater depth in Chapter 9).

In the symptom and side-effects, PHQ-8 and GAD-7 anchor analyses, differences were seen between the ICECAP measures and EQ-5D-3L. While the magnitude of change was similar, the pattern of change was different. In these analyses change in the ICECAP measures was larger in the group reporting a worsening of health state than in the group reporting an improvement. For the EQ-5D-3L the reverse was true: change was larger in the group reporting an improvement than in the group reporting a reduction.

CHAPTER 9: DISCUSSION

9.1. Chapter introduction

The discussion contains six sections. First, the results of the quantitative and qualitative research are summarised by drawing out themes running through the research. Second, the methodological strengths and weaknesses of the work are reflected upon. Third, the findings are discussed and placed in the context of existing research. Fourth, the implications of this research for practice and policy are assessed. Fifth, the key contributions of this research will be summarised. Finally, areas for future research are discussed. This section will identify gaps in the current knowledge base and suggest future research methodologies to provide the required evidence. A conclusion then draws the thesis as a whole to a close.

9.2. Summary of principal findings and research themes

During the conduct of the research and writing of this thesis it became evident that the results centred on a smaller number of broad themes. The principal findings of the research will be described using these themes, which run through the qualitative and quantitative work, through the two trials and through the cross-sectional and longitudinal results.

9.2.1. ICECAP measures are simple and feasible for use

The question of whether the ICECAP measures are feasible for use in a randomised controlled trial, which is designed to improve the physical health of participants, is central to the research presented in this thesis. If a measure does not meet the practical constraints of a trial environment then, however valid it is, it is unlikely to be used. Data from the methodological review, the qualitative research, the experience of recruiting trials to the quantitative research and the quantitative research itself allow an understanding of practical considerations when using a quality of life measure in research and whether the ICECAP measure meets these.

The qualitative research found that informants viewed the ICECAP-A measure as short, straightforward and easy to complete. This appears to be reflected in the completion rates of between 92% and 99% in the studies identified by the methodological review [117,172,173,175,177,180] and in the trials used in this research. These completion rates considered in conjunction with the results of the quantitative research indicate that the measures are feasible for use in trial-based research.

The frequency of use of a measure is a good indicator of its feasibility for use. The number of studies which have registered to use the ICECAP measures is increasing year on year. This

suggests that research groups see the benefits of using a broad measure of well-being that conceptualises quality of life through the lens of capability. In contrast, the experience of recruiting trials to participate in this research indicated that some researchers hold reservations about the measure. Two research groups declined to include an ICECAP measure in their trials due to concerns either with the content of the measure, or concerns over whether the measure would be responsive to the intervention being studied.

Concern about the content of the measure was an unexpected finding, which relates to the usability of the measure. A small number of researchers felt that the content of the ICECAP measures was inappropriate and could prove to be upsetting to participants. The qualitative research identified that these perceptions were motivated by the perceived potential for items being upsetting to people who had low levels of the dimensions being assessed, rather than being inappropriate or upsetting *per se* (e.g. asking someone about their attachment might be upsetting to someone who had little love and friendship in their life, but not to someone who had a strong friendship circle). The qualitative research also suggested that this was a minority opinion; the majority of informants found the content of the measure acceptable, with some showing notable enthusiasm for its content.

Taken as a whole the results indicate that the ICECAP measures are feasible for use in a randomised controlled trial designed to improve physical health. Concerns over the relevance and sensitivity of the content of the measure, held by some, were not reflected in the majority view of the qualitative informants or in the completion rates of the measures.

9.2.2. ICECAP measures go beyond health

Quantitative and qualitative data both indicate that the ICECAP-A and ICECAP-O measure a construct that is different from that assessed through health, physical functioning or health-related quality of life measures.

Using the comparative direct approach informants' conceptualisation of quality of life, their opinions of the ICECAP-A measure and whether the ICECAP-A captured *their* conceptualisation of quality of life was examined. In the early part of the qualitative interviews informants discussed a construct that was broader than health alone. A majority view was evident that while health was a strong influence on quality of life and well-being, other social influences such as family and enjoyment were also important. When considering the ICECAP-A measure, informants frequently observed that it was a broader measure than existing health focused measures. A number of informants, some spontaneously but often when prompted, noted that the measure captured the broad definition of quality of life and well-being they had previously described.

The quantitative results provide support for the informants' perceptions that the ICECAP measures reach beyond the measurement of health. First, moderate correlations were found with measures of health and physical functioning, which are suggestive of a non-perfect relationship between ICECAP capability scores and the scores of health measures. If the ICECAP measures were capturing solely health status, stronger correlations would be expected. These correlations indicate that determinants other than health are affecting scores in the ICECAP measures. Second, the measures were found to associate with the limited number of non-physical health measures included in the participant questionnaire packs.

Moderate correlations were found with vitality, social function and measures of psychological health. Third, the responsiveness analysis showed, in this population of high capability individuals, change in the ICECAP capability scores was smaller than change reported in the health anchor. This suggests that change in the health anchor was not a sole, or possibly even a major, determinant of change in ICECAP scores. Furthermore, the responsiveness analysis showed differences in the patterns of change between the ICECAP measures and the EQ-5D-3L comparator. For example, the ICECAP-A measure was more responsive to *reductions* in psychological health, whereas the EQ-5D-3L was more response to *improvements* in this construct.

9.2.3. A complement not a replacement for existing measures

The qualitative findings and quantitative results indicate that the ICECAP capability measures should be used in addition to existing health-related quality of life measures, rather than as a replacement. A strong theme running through the qualitative data was that informants viewed the ICECAP-A measure as capturing different information to that captured by existing health-related quality of life measures. Informants noted that ICECAP-A and EQ-5D-5L were measuring two different concepts: health, versus a more general conceptualisation of well-being. The broader assessment of quality of life was viewed as useful in addition to measures which maintained a focus on health or health-related quality of life.

The quantitative results provide support for the opinions of informants. The two factor solution to the factor analyses, found in both trials, indicated that the items of the EQ-5D-3L and the ICECAP measures are measuring two different concepts. EQ-5D-3L items, with the

exception of anxiety and depression item, loaded onto a factor; while the ICECAP items loaded onto another factor.

9.2.4. The importance of responsiveness

Data from both the qualitative and quantitative research indicate that the responsiveness of a measure is an important consideration for trialists and an area in need of more research with the ICECAP measures. An emerging theme from the qualitative research was how measures are selected for use in a randomised controlled trial. This emergent theme identified responsiveness as a central, if not the primary, consideration for trialists when choosing a measure. Informants discussed how such concerns had stopped them from using quality of life measures as an outcome and many were unsure whether it was possible to detect the effect of an intervention in existing quality of life measures. An example of the importance of responsiveness was reflected in the decision of one trial to decline inclusion of the ICECAP-A measure for the purposes of this research, citing concerns about the responsiveness of the measure to their intervention.

The responsiveness analysis found changes in ICECAP capability scores in response to changes in health. The changes in ICECAP scores were small, but consistently in the same direction (e.g. when health improved so did capability). The inferences that can be drawn from these results and the possible reasons for proportionately smaller changes than changes in the health anchor are discussed below. These results, in conjunction with the results indicating the importance of responsiveness to trialists and researchers, highlight the need for further research, which is discussed later in the chapter.

9.2.5. Capability as a new research area

The development of capability measures for use in health and social care research is in its infancy [100,115,159]. A small number of measures have been developed. Arguably the ICECAP measures are most advanced in this process, with a small number of validation studies already completed. The results from the methodological review, the qualitative research and the quantitative results reiterate that the development of capability measures is a new area of research and the validation of the ICECAP capability measures is in its early days. Results from each section of this thesis suggest the need for further validity research, which is discussed further below. Furthermore, the stage of validation has an implication for the conclusions that can be drawn from this thesis. The argument based approach [205] to validation means that validity is inferred as evidence is brought to bear. The weight of evidence informs the certainty of conclusions. Therefore in this early stage of validation cautious conclusions should be drawn.

9.3. Reflections on the strengths and limitations of the work

The research presented in this thesis has both methodological strengths and limitations which need to be considered in order to draw accurate conclusions from the findings. This is done through a reflection on the qualitative research and a discussion of the quantitative strengths and weaknesses.

9.3.1. Qualitative reflection

The potential limitations of the qualitative research were: the impact of the researcher upon the research; the characteristics of the informants interviewed; and the reaching of saturation point.

9.3.1.1. The influence of the researcher

A debate exists within the qualitative research community about whether the researcher is an objective figure who does not project values or opinions onto the research or whether it is impossible for the researcher to be neutral [270]. In this situation it is likely that I, as a researcher, had an (largely unquantifiable) effect on the data provided by informants and the interpretations drawn during the analysis. In this situation reflexivity or self-reflexivity is an important tool for interpreting work honestly and objectively [270,314]. This involves reflection on the ways, as the interviewer, I may have influenced the findings of the research and is a fundamental practice in good qualitative research. Three primary influences may have had an effect upon the results: my link with the work on the ICECAP-A measure; my relative lack of experience of qualitative research; and my position as a comparatively junior researcher to all but one of the informants.

As a researcher who works within the team that developed the ICECAP-O and ICECAP-A measures and whose thesis is focused on assessing the validity of these measures it is likely that my position may have influenced the research in two ways. First, a small number (five) of the informants were aware of the research group in which I worked, or had an understanding of the subject of my thesis. This knowledge may have resulted in these informants exercising greater caution when critiquing the measure or providing favourable

opinions in order to please. This was evident during an interview with one informant who showed perceptible embarrassment when describing his dislike of the ICECAP-A measure. For the other informants who had knowledge of my role, no such hindrance was observed, but findings may still have been influenced.

A second path through which my role may have influenced the results is an underlying bias affecting my conduct in the interview or my analysis of the results; in short I may have projected my opinion upon the research. Two steps were taken to reduce this effect. A topic guide was enlisted in conjunction with a cautious approach to ensure non-leading, open-ended prompts were used [278]. The transcripts of the early interviews were reviewed by supervisors to check that my interview style was suitable for the research and a conscious effort was made on my part to ensure that I developed and implemented an appropriate interview technique. Furthermore, a hierarchical coding structure was used in the analysis of results. This allowed the identification of themes within a formal structure. The coding structures and the output of the analysis were checked by supervisors at regular intervals. This coding structure and supervisor check will likely have reduced the impact of any personal bias upon the analysis process; although it is unlikely to have excluded it all together.

My confidence and ability as a qualitative researcher improved throughout the period of research. This was my first piece of qualitative research. Before starting, I completed a week long qualitative research course at the University of Bristol, which had a large practical component of interview practice. This gave me the basic skills required to conduct interviews and analyse their content, which were enhanced through the experience of the research

interviews[8]. By the end of the research interviews I had improved in four main areas: my ability to put the informant at ease in the interview (not least because I felt more at ease), my ability to identify themes and pertinent points during the interview and use appropriate prompts to increase the contextual richness of the data provided; my willingness to provide the informant ample time to reply to questions and not jump into every silence; and my ability to keep the informant on topic. The data provided in the early interviews may have been constrained due to my ability as a researcher. Therefore, data from these interviews may add less weight to the overall results, than data from the later interviews.

All except one informant was my senior, often with 20 years or more experience in research. This provided an interesting dynamic to the interviews: I adopted the role of a junior researcher [269]. In many instances the dynamic of the interview was close to a student and teacher relationship, where the teacher (informant) imparts knowledge upon the student (interviewer). This dynamic was logical, in most cases productive, and ensured a good working relationship during the interview. However, it did limit my ability to challenge inconsistencies. To explore inconsistencies in the opinions the informants provided I had to step out of the role as "student" and assume a more challenging position. When doing this the dynamic of the interview often quickly changed. In some cases, this led to a reduction in the data provided on that subject. For example, one informant defined quality of life in very broad terms and critiqued the ICECAP-A as being a broad measure. When challenged on this point she became defensive and quickly changed subject.

[8] Furthermore, both of my supervisors, who were experienced in the conduct of qualitative research and the analysis of qualitative data, provided ongoing advice and support throughout the work.

9.3.1.2. The characteristics of the informants

Informants were recruited to the research in their position as researchers involved in randomised controlled trials. Their levels of experience varied from senior professors or senior research nurses to early career researchers. A number of researchers had worked on numerous trials and some were viewed as leading national or international experts in their field. None of the informants had an expert knowledge of the theory of quality of life measurement; however all of them had experience of using quality of life measurement and many were experts in the practical issues of trial measurement and analysis of collected data. Despite this practical experience and expertise, even the researchers who were most experienced were, when asked, very clear about their perceived lack of knowledge or lack of professional research interest in the of quality of life measurement. Therefore, the views captured in the qualitative research can be considered as the views of health researchers rather than experts in the theory of quality of life. Effort was made at the start of the interviews to establish there were no wrong answers, to ensure that informants felt able to voice their opinions and it was emphasised that their opinions were of great value to the research [278]. Despite these efforts, it was evident that some informants were cautious about voicing their opinion and concerned about being "wrong". At times this meant that it was difficult to fully explore their views.

Informants were also recruited primarily from clinical areas. Only a small number of informants had experience in social care research or public health research. The ICECAP-A measure may be of particular use in a social care setting or in evaluating the impact of public health interventions [99]. While it should not be considered a weakness of a study seeking to assess the validity of a capability measure in randomised controlled trials, the results and

conclusions of this research may have been different had researchers from a social care setting been included in the research. This is discussed again in the future research section below.

9.3.1.3. The early achievement of data saturation

Data saturation [239] was reached quickly on the primary objective of the research: the face and content validity of the ICECAP-A measure . This is likely to be due to the relatively homogenous nature of the sample (all informants were involved in health-focused randomised controlled trials) and the focus of the research on a clearly defined question. No new themes were identified after the 14th interview. The relatively low number of informants should not be viewed as a weakness of the research. Data saturation suggests that continued recruitment of informants would have provided no additional information on the primary objective of research [315,316]. Data saturation was not, however, reached on the emergent theme of how measures are selected for use in a randomised controlled trial. This work provides some tentative conclusions on the how a measure is selected for inclusion in a trial and what is considered while choosing a measure. In light of the fact that data saturation was not reached these conclusions require further confirmatory research.

9.3.2. Strengths and weaknesses of the quantitative analysis

This thesis reports the first assessment of the construct validity of the ICECAP measures in a randomised controlled trial and the first analysis of the responsiveness of the measure using longitudinal data. In evaluating the strengths and weaknesses a number of aspects of the work are considered, which relate mainly to the randomised controlled trials and methodology used in the analysis.

9.3.2.1. The randomised controlled trials

The recruitment of the two randomised controlled trials to this research and the provision of data from these trials is a major strength of this research. The PastBP [279] and BEEP [280] trials were trials run by established research units in two different clinical areas. Bias was controlled for and data collection processes were rigorous. These trials are reflective research settings where it is likely that the ICECAP measures will be used in future.

The provision of data from two medium sized, randomised controlled trials with a sample size in both cases of over 500, makes this one of the largest assessments of the construct validity of the ICECAP measures to date. Both trials provided data from the majority of outcome measures completed by participants at baseline and follow-up. This provided a number of comparator measures for the baseline validity analysis and a wide choice of anchor measures for the longitudinal responsiveness analysis. Therefore, this research provides a large amount of data, substantially adding to the existing validity portfolio, and can be used to inform future validity research.

The provision of this data from two trials (rather than one) in different clinical areas has increased the generalisability and applicability of the research. However, as discussed in Chapter 3, the psychometric properties of a measure are context specific [200,205,206]. Caution should be exercised in applying the results outside of these research settings. As is the case with the majority of randomised controlled trials, the inclusion and exclusion criteria of the PastBP and BEEP trials resulted in a specific sub-section of the general population being included in the trial. The PastBP inclusion criteria were those who had suffered a stroke or TIA, had a systolic blood pressure over 125mmHg and were not taking three or

more anti-hypertensive drugs. The BEEP trial randomised those who had consulted GP due to knee pain in the last 12 months or were being referred to a specialist due to knee pain. The result of such selection criteria is to select a group of participants that is different to the general population on a number of key characteristics [119]. Therefore, caution needs to be exercised when generalising these results to either the general population or other trial populations. The characteristics of these populations are presented in detail in this thesis, which it is hoped will assist researchers when deciding whether it is appropriate to generalise the results to other populations.

Two participant characteristics are of particular note. Firstly, the EQ-5D-3L index scores of 0.63 in the BEEP trial and 0.72 in the PastBP trial indicate that while the participants do have health problems they are not, on average, in a debilitative health state. Different values may be found in other populations with the same health issues or with populations with different health issues. For example, a late stage cancer patient might be expected to be in worse health. Second, the ICECAP tariff scores of patients in both trials were comparable to or higher than the general population. It is not possible to comment if the capability of these participants was unusually high for a trial population or whether trial populations generally exhibit high capability levels. Either way the quantitative results of this thesis need to be considered in light of these high capability scores.

The PastBP trials used a pharmaceutical intervention to reduce a person's blood pressure, while the BEEP trial used a physical intervention to increase a person's recovery from knee pain. To what extent these interventions would change the capability of a person and to what extent this affects a psychometric assessment is an important consideration when interpreting the results of this research. It was the hope of the supervisory team at the outset of the

research that a trial where patients had considerable disability and had a notable improvement in health or physical functioning, such as a hip operation, or a trial of a social care intervention, such as the maintenance of independent living, would be recruited to the research. No trials of this type that would provide data within the timespan of the PhD were found. Large changes in the capability of the participants in the recruited trials were not expected. This was reflected in informal conversations amongst the supervisory team, as well as with the trial teams during the continuation of the research. This was particularly so for the PastBP, where high blood pressure is largely non-symptomatic until a cardiac event occurs.

To what extent this matters to a psychometric analysis, especially the analysis of responsiveness needs to be understood clearly. This analysis did not use the randomisation of participants in this analysis. So no analysis of those receiving intervention versus placebo was completed. Rather, the responsiveness analysis used anchors to define groups which had changed by some degree on some measure and then ICECAP scores were analysed in these groups. Therefore, the analysis of responsiveness does not rely on the success of the intervention; rather a level of change in the population needs to have occurred. Analyses presented in Chapter 8 indicate that the majority of participants reported small changes in the EQ-5D-3L and ICECAP scores occurred in these populations. Therefore, the responsiveness analysis was completed in a population showing low levels of health and capability change.

9.3.2.2. The methods

The methodology used for this analysis has a number of strengths. It has utilised best practice from established and well reported research techniques in the field of quality of life measure validation [185,186]. Firstly it used a scientifically rigorous process of hypothesis formation

and testing. Hypotheses were formed through a pre-planned process [200]. Three researchers independently formed hypotheses. These were then compiled and confirmed for final testing when all researchers agreed on the direction of the association. Where post-hoc or exploratory analysis has been completed (which did not involve the use of stated hypotheses) this is clearly highlighted in the text and caution has been exercised when interpreting the results.

The methodology and statistical analysis used to test hypotheses is appropriate [138,184] and has been used consistently in the testing of the psychometric properties of patient-reported outcome measures [116,117,287,317]. Known groups, convergent and divergent validation has allowed this research to show what measures the ICECAP O and ICECAP-A associates with and what the discriminative ability of these measures are. Anchor based analyses allowed the assessment of change in ICECAP scores when a change in the health anchor occurred. This has the benefit of allowing the results to be easily interpreted in light of past and future research.

This research methodology also has some limitations which should be noted, namely, the predominance of health measures alone being available for use as comparators and anchors. In this analysis the validity and responsiveness of the ICECAP measures were tested using data from measures already included in the randomised controlled trials. Hypotheses were formed and correlations and associations were assessed against these measures. As is shown in the results chapters and discussed here, this provided informative and novel information to contribute to the validity and responsiveness portfolio of the ICECAP measures. However, the predominance of health, physical functioning and health-related quality of life measures and the lack of measures of social well-being (which are rarely included in trial questionnaire

packs), means that the scope of this assessment of validity and responsiveness has its limitations.

The ICECAP measures are an assessment of a broad conceptualisation of well-being, capability. Grewal et al [109] identify health as one of six broad factors that impact a person's capability, including having activities and things to do, home and surroundings, family and other relationships, standard of living and wealth and religion or faith; while Al-Janabi et al [115] identify five conceptual attributes that determine an individual's capability, each of which is assessed in the ICECAP-A measure via one item. This forms the empirical basis upon which the ICECAP-O and ICECAP-A measures were developed. In the research for the development of both ICECAP measures health is acknowledged as one of a number of influences, albeit important, upon the capability of a person. This was reflected in the development of, and ultimately in the final versions of, the ICECAP measures. The data made available by the trials for use as comparators and anchors in this analysis predominantly assessed the health of the individual.

In Chapter 3 the use of a nomological network was defined and discussed. The idea that a number of stated hypotheses about how the measure under investigation was expected to associate with other measures should be tested is a widely accepted method of construct validation [203]. The data available from these trials meant that hypotheses on how the measure associated with measures or indicants of health and physical functioning could be tested; while associations with the other determinants of capability went under assessed. For example, in both ICECAP measures the ability to achieve and the ability to feel secure or stable are important constructs. Data from measures which assessed such constructs, or similar constructs, were not available. Therefore, no hypotheses could be formed or tested

which provided information on the association of the ICECAP measures with measures of achievement or security. Consequently, a complete nomological network could not be tested. This challenge also applied to the responsiveness analysis where the change in the measure was assessed against change in health and physical functioning measures, but not against changes in other influences on capability.

These limitations in the data mean that this analysis seeks to validate and assess the responsiveness of two measures which have theoretical grounding in the capability approach by using outcome measures which are largely couched in the health economics extra-welfarist school of thought [7]. As discussed in chapter one, the practical exposition of the extra-welfarist approach has a limited evaluative space of health or health-related quality of life [7,318]. While not all of the measures included in the PastBP and BEEP trial were included with the aim of informing the economic analysis of the trial, the majority would be suitable for use either in a cost-effectiveness analysis or in the case of the EQ-5D-3L and SF-36 (that can be transformed into the SF-6D preference based measure [153]), a standard cost-utility analysis [88,89,319].

This predominance of health-focused outcome measures is not exclusive to the trials included in this research. The majority of randomised controlled trials in health research will focus on health outcomes. If the ICECAP measures are to be used in randomised controlled trials, then their psychometric properties must be validated in such a setting. So while the scope of this analysis is truncated, this should not be considered a limitation which is exclusive to this research or a reason for not continuing validation research into the ICECAP measures in randomised controlled trials. It is a challenge that this research faced and one that future evaluations of capability measures in randomised controlled trials will also face: how to

validate a capability measure using predominantly health outcome measures. This point is discussed further in the future research section.

9.4. Discussion of principal findings

This section will discuss the findings of the qualitative and quantitative research in the context of existing research, much of which is presented in the methodological review in Chapter 2.

9.4.1. Content and face validity

The need for triangulation of qualitative research that assesses the face and content validity of a measure is discussed in Chapter 3 [242,244]. Recently published guidelines for assessing content validity highlight the need for comparison between research completed with both experts and public or patients in reaching a conclusion on the validity of the content of a measure [233]. This allows the integrity of results to be checked and, potentially, inferences extended [244]. The qualitative "think-aloud" research by Al-Janabi et al [181] allows for triangulation with research using the general public as informants[9]. Comparison can be made on five points: ease of completion, understanding of capability wording, breadth of the questions, whether the questions are upsetting and the relevance of the Achievement item.

The results show expert informants viewed the ICECAP-A as a short, easy to complete measure and suitable for use in a trial environment. Particular emphasis was placed by informants upon how well the measure fitted with the time constraints that exist in such research. This confirms the findings from research with the general public, which found the measures straightforward, unsurprising and possible to complete with a low level of error

[9] A qualitative "think aloud" study by Horwood was identified through the methodological review. [182]. This study used the ICECAP-O and therefore is not used here

[181]. Together, these results indicate that the ICECAP-A measure is an easy to complete measure (and quantitative results, discussed above further confirm this).

The use of the phrase "I can" and "I am able" (to indicate capability) was discussed by informants in the qualitative research both in this thesis and in the "think aloud" research with the general public. While the majority of expert informants demonstrated an understanding of the capability wording that was broadly in line with capability theory, a number of informants expressed concern about the phrasing of the questions. Expert informants felt that the wording may be confusing for participants completing the questionnaire. The "think aloud" work [181] showed that some informants interpreted the questions as assessing functioning, or they struggled with the use of the phrase "I can", sometimes interpreting it as a question of worthiness. However, not all general public informants struggled with this phrase and many reached an understanding which was correct. Together these data indicate that participants may experience some problems interpreting the capability wording of the questionnaire and draws attention to the importance of the phrasing of capability questions in future questionnaires.

A small minority of the expert informants objected to the content of the measure on the basis that it may prove upsetting for informants. Those who did object were considering the situation where a respondent had low levels of capability and answering questions about this may prove upsetting. No indication of this was found in the research with the general public [181]. This, in conjunction with the majority of expert informants who either felt the measure was appropriate or did not comment on this issue, suggests that this issue may not be widespread.

Contrasting findings were found between this research and the "think aloud" research on the suitability of the breadth of the ICECAP-A items. Al-Janabi et al [181] found that the breadth of the items allowed 'participants to bring a large range of factors, which influenced their well-being into their answers'$_{(p.120)}$. The findings from the research with experts indicated that, with particular reference to the Attachment item, informants had concerns about items assessing more than one construct. The example was given of a person having love, but not friendship and support, and the question was raised how someone would then answer the Attachment item.

Informants from both pieces of research raised the question of whether the Achievement item was relevant for older people (the ICECAP-A being applicable for all adults). Some of the older general public informants questioned its relevance to people at their (later) stage of life, making the point that for many at their stage of life it is about maintaining what they have, rather than achieving more [181]. The expert informants went further, firstly by questioning whether it would be appropriate for older respondents, but also whether the top level is achievable for anybody. It should be noted here that this concern was not borne out in the quantitative results where large percentages of both the older PastBP and younger BEEP trial populations reported full capability for the Achievement item. However, the results of both qualitative studies indicate that further consideration of this item of the ICECAP-A measure may be required.

As discussed above, the ICECAP-A measure was viewed as a complement rather than a replacement for existing health-related quality of life measures. Informants discussed the use of the ICECAP measures alongside other measures. Health economist informants discussed using it alongside preference or value based measures, namely the EQ-5D-3L. It is usual for

trials to include more than one health-related quality of life measure in their questionnaire packs. It is recognised that inclusion of multiple outcomes for assessment of the clinical effectiveness of the trial is acceptable as long as the primary and secondary outcomes are specified *a priori* in the trial protocol [125]. The use of multiple outcomes may increase the depth of the information captured. For the purposes of health economic analysis alongside a randomised controlled trial, the norm is that one preference or value weighted measure is included in the questionnaire pack and the results of this measure is used to form QALYs which inform an economic analysis. The traditional choice has been for a single measure to be used to form a QALY. Therefore, the potential of use of the ICECAP measures in addition to existing measures may be a possibility for researchers assessing the effectiveness of a treatment. However, in light of current and accepted norms for health economic measurement alongside randomised controlled trials this may not be an option, in spite of health economist informants raising this possibility.

In summary, a number of similar themes were identified in both the qualitative research from this thesis and past research by Al-Janabi et al [181]. Triangulation has allowed results to be compared and contrasted under these themes. Considered as a whole and together, the qualitative research into the ICECAP-A measure provides encouraging evidence of content and face validity.

9.4.2. Construct validity

9.4.2.1. Meaning of results in context of past research

The results from the validation studies using the PastBP and BEEP trials can be compared with findings of the literature included in the methodological review. The associations of

ICECAP measures found in these trials largely confirm the associations found in previous studies that were reviewed and the majority of stated hypotheses were confirmed by results.

The response profile of the ICECAP-O in the PastBP trial was similar to findings by other studies with both the general public [117,178] and patient populations [176,180]. Out of the five ICECAP-O items the Attachment and Control items received the highest proportion of respondents reporting full capability, which is reflective of past research. The mean ICECAP-O tariff score was at the top of the range of scores seen in past research [116,175,176,178,179] and the ICECAP-A tariff scores was a little higher to the only existing value in the general population [116].

Moderate correlations with measures of self-reported health were found, which are suggestive of a non-perfect relationship between the ICECAP measures and health. This is similar to findings from past research. The correlation of the ICECAP-O with the EQ-5D-3L was slightly higher than previously found [173–176]; while the correlation for the ICECAP-A was also moderate.

The ability of a measure to discriminate between individuals in different health or capability states is an important characteristic of valid measures to be used in health research. The known groups analysis using the Modified Rankin Scale, the symptoms and side-effects or co-morbidities questionnaires, the GAD-7 and the PHQ-8 provides encouraging indications that the ICECAP measures can differentiate participants with different levels of psychological and physical health. This reflects past findings which showed significant differences in ICECAP scores between those reporting different levels of general health [177] and psychological health [172,180].

The item-by-item analysis showed some additional unexpected associations. The closer than expected association of the ICECAP item of Attachment with measures of health showed similarities to results found by Al-Janabi et al [116] and Ratcliffe et al [176]. Using factor analysis, Davis et al [173] have previously suggested that the ICECAP-O and EQ-5D-3L are measuring two separate, but correlated factors of "physical functioning" and "psychosocial well-being". The factor analyses in this research using the ICECAP-O confirm this two factor solution found by Davis et al. The first factor analysis using the ICECAP-A, which shows a very similar two factor solution, is also presented herein.

The availability and use of data from the psychological scales of the SF-36 with the ICECAP-O and the GAD-7 and the PHQ-8 with the ICECAP-A, means that this is the most comprehensive assessment of the associations between the ICECAP measures and measures of psychological health to date. Past research using the Hospital Anxiety and Depression Scale [172] and the Herth Hope Index [176] provided mixed and incomplete results on the associations with psychological health. Our analysis shows consistent results confirming the hypothesised association of the ICECAP measures with measures of psychological health. Both ICECAP measures show moderate correlations with measures of psychological health and the known group analyses show the both measures are able to discriminate between participants in different levels of psychological health.

9.4.2.2. Conclusions to be drawn

As discussed at a number of points in this thesis, validity is context specific and this is the first assessment of its kind in this context. The argument based approach to validation, proposed by Kane [205], explains that evidence needs to be brought to bear and that the

certainty of conclusions increase as more and more evidence is provided. As the first assessment of validity in this setting, the conclusions drawn from this work need to be relatively cautious. Previous research from other settings can assist in the interpretation of the results, but do not allow firm conclusions to be drawn.

The results presented in this thesis provide encouraging initial evidence of the validity of the ICECAP measures in a randomised controlled trial setting. Associations which were expected *a priori* were found for both measures. The results are suggestive of a measure that is capturing a conceptualisation of quality of life that is broader than health or health-based assessments of quality of life. This broad measure appears to be able to discriminate between groups in different states of health or physical functioning. Further research is required to confirm this finding, and the potential nature of this research, is discussed below.

This finding will be of interest to health economists. This provides an early indication that capability may be accurately measured through the use of value weighted, patient reported outcome measures in a randomised controlled trial. As discussed in Chapter 1 the extra-welfarist approach has been criticised for restricting the scope of its evaluative space. The finding that capability may be accurately measured in a trial environment, alongside which the majority of health economic analyses currently take place, should prompt health economists to question again the close reliance on measures of health.

This finding will also be of interest to trialists and health researchers. The ICECAP measures are currently being used in a number of research projects. These findings will allow trialists and researchers to have greater confidence that the measure offers a broader conceptualisation of well-being and the inferences that we can draw from it are accurate and sound. Research

which previously showed encouraging evidence of validity in the general public and patients, has been extended to a interventional health research setting. Conclusions of their trial participants' capability can now be measured with more confidence.

Capability researchers and those interested in the application of the capability approach in health will likely be interested in the findings. The findings indicate that Sugden et al's [54] question over whether the approach could be operationalised has been (largely) answered in the area of health, through findings that show strong indications of capability measure validity. It also provides an indication that development of capability measures at the methodologically driven research end of Robeyns [111] continuum (rather than at the theoretical end) may be the most promising route to operationalizing capability measures in health.

9.4.3. Responsiveness

This is the first analysis to assess the responsiveness of the ICECAP-A or ICECAP-O measures using longitudinal data. The results from this analysis provide evidence for the responsiveness of the ICECAP measures. Results show small changes in ICECAP-O and ICECAP-A tariff scores in response to changes in health. The majority of SRMs and effect sizes for these changes are small to moderate, indicating that the proportion of change to standard deviation of baseline score or change score was small.

Analyses using the non-weighted ICECAP scores, which were calculated by summing the raw item scores without applying the tariff values, indicate that the small levels of change in the value tariff scores may in part be due to the values of this tariff. The use of the EQ-5D-3L as a comparator measure allows further context to be added to the results. Comparison of non-

weighted scores and value weighted ICECAP tariff scores with the non-weighted and preference weighted EQ-5D-3L scores allows comment on the change in ICECAP scores in the context of another value weighted measure. As has been stated above, this comparison is not designed to assess which measure is "best", rather to contextualise the results.

These results were found in two populations which had high capability, reasonable physical health and good psychological health at baseline and follow-up. Furthermore, these results were found in trials of two interventions that were attempting to improve specific areas of physical health. The context in which these results were found is an important consideration when interpreting the results. Different interventions will attempt to improve different aspects of physical and psychological health, or sometimes aspects that stretch beyond health. Changes in different health areas will likely have differing impacts upon capability. For example it is unlikely that the changes in capability that are brought about by, for example, a hip replacement, cancer treatment and impotence treatment will be the same. It seems likely that some treatments will have a larger influence on capability than others.

9.4.3.1. Non-weighted versus value weighted scores

A comparison between the non-weighted ICECAP change scores and the value weighted ICECAP tariff change scores shows that change as a percentage of possible change is smaller for the value weighted tariff scores than for the non-value weighted scores. This trend was found in both the ICECAP-O and ICECAP-A measures. This indicates that when the value tariff was applied to the non-weighted scores the magnitude of change was reduced.

The value tariffs for the ICECAP-O and ICECAP-A were calculated using best-worst scaling [118,166]. there are differences in the value attached to change between the different item

levels. Significant value is attached to change between "none" and "a little", changes between "a little" and "a lot" hold moderate value and little value is attached to changes between "a lot" and "all". This is shown in Figure 41, which presents data taken from Coast et al's [118] work on the valuation of the ICECAP-O. As can be seen a change between "none" and "a little" has greater value, and results in greater change in the overall ICECAP tariff score, than a change between "a lot" and "all". The exception to this rule was the Security item which showed similar differences across all levels. This is not the case for the non-weighted analysis where change between all of the levels held the same weight. A similar pattern is seen for the best-worst scaling of the ICECAP-A measure [166].

Figure 41: Tariff values of ICECAP-O measure, taken from Coast et al (2008)

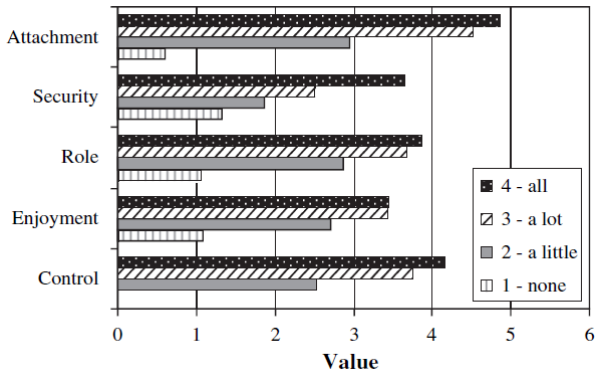


Table 2
Terminology for attribute levels, and rescaled values, such that the absence of capability, state 11111, is equal to zero, and full capability, state 44444, is equal to one ($n = 255$)

Attribute	Value
Attachment	
I can have all of the love and friendship that I want	0.2535
I can have a lot of the love and friendship that I want	0.2325
I can have a little of the love and friendship that I want	0.1340
I cannot have any of the love and friendship that I want	−0.0128
Security	
I can think about the future without any concern	0.1788
I can think about the future with only a little concern	0.1071
I can only think about the future with some concern	0.0661
I can only think about the future with a lot of concern	0.0321
Role	
I am able to do all of the things that make me feel valued	0.1923
I am able to do many of the things that make me feel valued	0.1793
I am able to do a few of the things that make me feel valued	0.1296
I am unable to do any of the things that make me feel valued	0.0151
Enjoyment	
I can have all of the enjoyment and pleasure that I want	0.1660
I can have a lot of the enjoyment and pleasure that I want	0.1643
I can have a little of the enjoyment and pleasure that I want	0.1185
I cannot have any of the enjoyment and pleasure that I want	0.0168
Control	
I am able to be completely independent	0.2094
I am able to be independent in many things	0.1848
I am able to be independent in a few things	0.1076
I am unable to be at all independent	−0.0512

This responsiveness analysis was completed in a high capability population. The extensive item by item analysis shows that the majority of change in this population occurred by respondents switching answers between the top two levels of both measures. Therefore, the

majority of change occurred at the top of the measure. When applying the value tariff these changes are of less value, and contribute less weight to the overall tariff score, than change at the bottom of the measure. Consequently these "top end" changes held less weight in the value weighted tariff score than they did in the non-value weighted score. This accounts for the reduction in the size of change in the value tariff in comparison to the non-weighted scores.

This distinction is important. The results show that the responsiveness of the descriptive systems of the ICECAP measures is reduced when the tariff is applied because changes at the top end are not strongly valued. Therefore, in high capability populations not experiencing large changes in capability the ICECAP tariff scores could be seen as not very responsive, even though the descriptive system of the ICECAP measure has detected change. In populations with low levels of capability this may not be the case. Here, the value weighted ICECAP tariff weightings may better reflect changes at the bottom end of the measures, because these changes hold greater value.

9.4.3.2. EQ-5D-3L reference measure

The use of the EQ-5D-3L as a reference measure showed that the size of change, the effect sizes and standardised response measures were similar for the non-weighted EQ-5D-3L scores as they were for the non-weighted ICECAP-O and ICECAP-A scores. Change as a percentage of possible change in the two measures was similar and both measures showed small to medium effect sizes and SRMs on the majority of anchor analyses. While the size of change and the signal to noise ratios were similar, the pattern of change was, at times, different. The ICECAP-O showed greater change in those whose symptoms and side effects

had worsened than improved. The ICECAP-A showed greater change in those whose psychological health had worsened, than in those in which it had improved. In both instances the reverse is true of the EQ-5D-3L.

While there were differences between the two measure the results of this research indicate that the responsiveness of the non-weighted ICECAP-O and ICECAP-A capability measures was similar to the non-weighted EQ-5D-3L health-related quality of life measure. This means that in response to changes in health anchors, the descriptive systems of these two measures, which are measuring different constructs, seem to show similar levels of change.

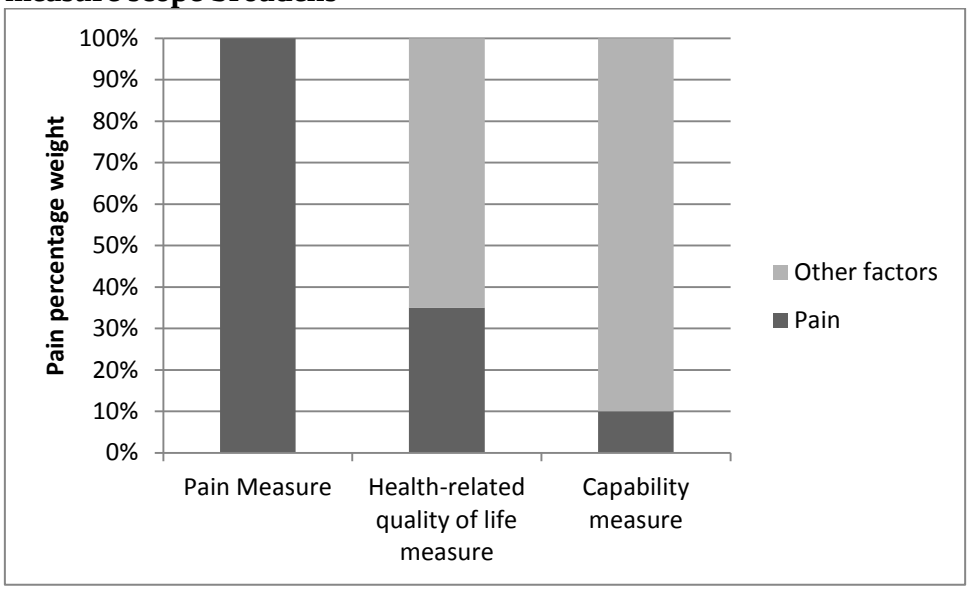
There were greater differences between the preference weighted EQ-5D-3L index score and the value weighted ICECAP tariff scores than there were in the non-weighted analyses. As is discussed in the previous section there were differences between the value weighted and non-weighted ICECAP scores. This trend, of the value (or preference) based scoring system reducing the magnitude of change, was less evident for the EQ-5D-3L index score. However, these differences were, in most cases, still small and effect sizes and SRMs were small or medium in both value weighted measures.

9.4.3.3. A non-perfect relationship between health and capability

As with the assessment of cross-sectional validity (where moderate correlations were suggestive of a non-perfect relationship) the results of this responsiveness analysis indicate a non-perfect relationship between change in health and change in capability. Correlations between the anchor change scores and ICECAP change scores were weak and the change in ICECAP value tariffs were smaller than change in the anchors (judged as a percentage of possible change). Considerations of the reasons for this non-perfect relationship is important.

A change in health represents a change in one of a number of factors that affect a person's capability. The descriptive systems of the ICECAP measures are constructed based on the broad range of factors identified by Grewal et al [109] for the ICECAP-O and Al-Janabi [115] for the ICECAP-A. Therefore, the impact that a change in health has upon the descriptive system of the ICECAP may be small. The following hypothetical example highlights this (Figure 42). Take three measures: an aspect specific measure of pain, a health-related quality of life measure and a capability measure. For the aspect specific measure, pain is the only determinant of scores on that measure. Scores on the health-related quality of life measure are determined by a number of health related factors, with pain being an important determinant. The capability measure scores are determined by a number of health and non-health related factors. Therefore, pain may have a considerably smaller impact on capability scores than on the scores of the other two measures.

Figure 42: An illustrative example of the potential of impact of single aspect of health as measure scope broadens



While this description does seem the most convincing justification for a non-perfect relationship between change in health and change in ICECAP scores, a number of other reasons also need to be considered. First, people may adapt to changes in health, which results in the changes not being reflected in how they rate their capability. Sen [48,53,320] has described at length the potential for adaptation of health and desires, in part as a justification for the use capability as a measure. If people are adapting to their health state so that changes are not reflected in self-reported capabilities, then this justification for the use of capabilities may not hold in patient-reported outcome health research.

Second, the changes in capability brought about by changes in health could be time lagged and not captured in the timeframe of the trials. It may be that, for example, only after many years or months reduced attachment that reality overtakes memories when an individual describes their love and friendship.

Third, changes in health may change the basic capabilities of an individual, such as the ability to live without disease and the ability to have bodily integrity, while the higher order capabilities assessed by the ICECAP measures may not be affected. Sen [53,76] proposes that in some situations the assessment of basic capabilities may be appropriate. Sen normally states this with reference to poverty. However, if the descriptive system of the ICECAP measures, which assess higher order capabilities are not responsive, then this may raise the question of whether more basic capabilities are appropriate for assessment in a health research setting.

Two further explanatory factors of the relationship between health and capability scores, which are specific to this research are the high ICECAP scores at baseline found in both the PastBP and BEEP samples and the possible selection of a sample that is different from the general population. The mean ICECAP capability scores at baseline are higher than any of the mean scores found for these measures by the methodological review. The ICECAP-O tariff score of 0.86 is higher than mean values found in the general public of 0.814 [118] and 0.832 [177] and in patient populations of 0.779 [176] and 0.753 [178]. The ICECAP-A score of 0.89 is higher than values found in the general public of 0.832 [116] and these are the first values for a patient population. While the ICECAP scores were above average, the level of health in the trial populations (measured by the EQ-5D-3L) was either, in the case of PastBP, slightly below the average for the age group (0.75 versus 0.78) or, in the case of BEEP, considerably below the average for the age group (0.64 versus 0.8). Therefore the populations used in this analysis have better than average capability, but lower than average health in comparison to the general population.

Most trials have selective inclusion and exclusion criteria which results in a population with different characteristics than the general population [119]. The inclusion criteria for the PastBP and BEEP trials have resulted in a population with high capability and below average health; in these individuals capability is high despite their reduced health. This may be suggestive of a population in which the factors other than health have a large impact on their capability. Sen [53] draws the distinction between agency and well-being and factors identified by Grewal [109] and Al-Janabi [115] find health to be one of a number of factors influencing capability. It is possible to suppose that participants in the PastBP and BEEP trials have a level of independence which allows them to participate in a trial, or a support structure which does. These individuals may also be able to participate in activities, such as clinical trials, which make them feel valued. Therefore it may be that this is a population where agency, rather than well-being or health, is having a larger influence on capability and changes in health which occurred in these trial participants have had a disproportionately small impact on their capability.

A second potential influence of the high ICECAP scores and low health scores upon the responsiveness results is a methodological one. If the ICECAP scores are high at baseline and the scores of the health measures are low, when a participant's state improves there is more scope for improvement in the health measures than in the ICECAP scores. The results from the BEEP trial indicates this may be the case: a decrease in health score results in a decrease in ICECAP scores, but an increase in health score results in a proportionately much smaller increase in ICECAP scores. This is suggestive of a ceiling effect, where the sensitivity is lost at the top of the measure due to a clustering of responses towards the maximum score of the

measure. Ceiling effects can be fully assessed through comparison with another measure of the same construct, which is not possible here.

In summary the longitudinal responsiveness analysis provides an initial indication of the responsiveness of the measure. Interpreting these results has been challenging due to the challenges involved in this research. A number of potential explanations are proposed which deserve further research before a conclusion can be made with any form of certainty. This is discussed in the directions of future research section below.

9.5. Implications of research for practice and policy

The research presented in this thesis is largely methodological. It seeks to assess whether the ICECAP measures can be used with confidence in a research setting. Therefore its primary contribution is to health and social care research practice. However, in light of recent guidance by NICE and SCIE on the use of broad outcome measures, this methodological research also has some important implications for decision making and policy.

The methodological review and primary research presented in this thesis have provided evidence that the ICECAP measures are feasible and valid for use in health research and randomised controlled trials. Early evidence is presented indicating that the ICECAP measures respond to changes in health and the magnitude of change is in many cases similar to changes seen in the frequently used EQ-5D-3L. Furthermore, the qualitative and quantitative results from this research show that the ICECAP-A and ICECAP-O are measuring a construct that is broader than health.

These findings should allow health and social care research groups to use the ICECAP measures in health and social care research settings with greater confidence than has previously been the case, and this may lead to more frequent use. In doing so research groups may be able to fully capture the impact of an intervention on broader well-being, rather than constraining their analysis to the effect on health. This should allow a more accurate conceptualisation of effects of interventions that reach beyond health.

While the evidence of validity and responsiveness is encouraging, research into these measures is still at an early stage and has only been assessed in two specific contexts. Researchers considering use of these measures are cautioned against the assumption that the measures have been validated for use in all areas of health and social care research. They have not. In many research areas it may be suitable for the researcher to complete a validity and/or responsiveness analyses of the ICECAP measures in parallel to their use in an effectiveness analysis. This is particularly the case in research in clinical areas, or with populations, which do not show similarities to the PastBP and BEEP trials or the studies included in the methodological review. For example, there is no known example of the ICECAP measures being validated in a cancer population. The first research group to do so should complete a validity analysis of the measure alongside their research.

The results of this thesis have some important policy implications. If the effects of a treatment or intervention reach beyond health, then the use of health focused outcome measures currently used will under represent the effects of the treatment. The use of a measure that accurately captures the full range of benefits of a treatment will provide a better estimation of the impact and value of an intervention. Therefore the use of broader measures of well-being, such as the ICECAP measures, in a decision making context will likely lead to

different decisions being taken, due to what is being measured and valued. NICE and SCIE have highlighted a need for the use broad preference based measures in social care research [169,170]. The ICECAP measures have been identified by NICE and SCIE as two measures through which this may be achieved. However, both decision making bodies have cautioned that validation efforts are at an early stage and currently recommend the use of these measures in tandem with other broad measures such as ASCOT, in order to increase understanding of these measures.

The ICECAP measures appear to fulfil the NICE and SCIE criteria that measures used in social care should be broad, preference-based assessments of well-being. Furthermore, the strong indications from this research (which is the largest study of validity and responsiveness in a health research setting to date) that the ICECAP measures can validly assess capability and are responsive to changes in a health, in a health research setting, should increase confidence in the use of ICECAP measure outcomes in a decision making process and go some way to allaying concerns, expressed by NICE and SCIE, about the lack of validity evidence.

It should be noted that there is a difference between the areas in which NICE and SCIE have recommended the use of broader measures of well-being and the areas in which the ICECAP measures have been validated. The area of social care has been highlighted by NICE and SCIE as a suitable place for the ICECAP measures to be used, while the majority of research to date has validated these measures in a health setting[10]. This has two implications. First, it highlights the need for research into the validity and responsiveness in a social care setting.

[10] Please note that NICE and SCIE guidance was provided in 2013 and therefore was not available at the point when trials were being recruited to the quantitative research and informants recruited to the qualitative research.

However, logically, a degree of confidence should be held that the measures will be valid and responsive in a social care setting, where many of the interventions are designed to have benefits outside of health. Second, the ability to accurately measure in a health research setting a broader construct than health, may enable NICE and SCIE to consider whether it is appropriate to extend social care research outcomes guidance to health interventions. While a broader debate about what outcomes from health care a society should value is needed before such a decision could be taken, a research portfolio is starting to emerge which suggests that if broader outcomes were to be valued, they could be accurately measured.

9.6. Contributions of this research

The qualitative and quantitative research presented in this thesis has made four primary contributions. These are discussed below with reference to the existing research base and with comment on how they may inform practice in the future.

9.6.1. New qualitative methodology for assessing content and face validity

The PROMIS working group recently noted the inconsistency of content validation efforts and called for improvements to the methodology used [233]. Meticulously documented and transparent analysis, which is grounded in the data, is needed [266]. The comparative direct approach was developed in this thesis in response to the challenges faced in validating the ICECAP-A measure and may be a method through which requirements outlined by the PROMIS working group could be accomplished. The approach encourages participants first to discuss and describe a theoretical concept, and then comment on how the measure under

consideration captures their description of the concept. This approach could prove a useful methodological framework for validating the content of patient reported outcome measures going forward as many of the strengths of this methodology are generic. The ability to gain a fuller understanding of the informant's conceptualisation of a non-tangible concept, such as quality of life, well-being, happiness or capability, before assessing their views upon a measure of that construct, proved extremely useful in this research and could be beneficial to other research.

9.6.2. The first analysis of content validity using research professionals

Triangulation of qualitative research using both research professionals and patients or the general public as informants allows a fuller understanding than either on its own [233]. This thesis presents the first analysis of the content and face validity of the ICECAP-A measure. This allows triangulation with existing research (see below), which extends the inferences drawn from both this study and the "think aloud" work by Al-Janabi et al [114]. In addition to an assessment of content validity this qualitative work provides the first in depth analysis of the acceptability of a capability-based patient reported outcome measure amongst health research professionals working in randomised controlled trials. Interviews with researchers with different professional roles allowed investigation of whether the phrasing of the questions was thought to be understandable and whether the content of the measure was considered by these health professionals to relate to their conceptualisation of quality of life and to be appropriate and suitable for use with patients in their trials.

9.6.3. The first analysis of construct validity in a randomised controlled trial setting

A growing body of research assessing the validity of the ICECAP measures in patient and general public populations exists. Validity is a context and population specific quality. Evidence of validity of a measure in a general population survey, does allow the assumption that the measure is valid in a randomised control trial. This study is the first analysis of the validity of the ICECAP measures in the context of randomised controlled trial. *A priori* hypotheses were formed through a rigorous process of hypothesis formation and tested using appropriate measures of association. This analysis provides early evidence that the ICECAP measures are valid for use in an intervention health research setting and in particular the first evidence for the validity of the ICECAP-A measure outside of a general population.

9.6.4. The first analysis of responsiveness

Responsiveness is an important psychometric property and this qualitative research shows that it is an important consideration for research professionals when choosing a measure for use. This research provides the first analysis of the responsiveness of the ICECAP measure using longitudinal data. Results are presented from trial populations that showed reasonable levels of health and capability levels similar or higher than values previously recorded in the general population. Results indicate changes in the ICECAP tariff occur in response to changes in underlying health. Comparison between non-value weighted scores and value weighted tariff and between the ICECAP scores and EQ-5D-3L scores allowed greater interpretation of the data.

9.7. Directions of future research

This is the first research to assess the validity of the ICECAP measures in a randomised controlled trial, the first qualitative assessment of the content and face validity of the measures and one of the first validation studies *per se*. In addition to the 3 areas for future research area identified below, there is a general need for greater study into the validity of these measures. Kane proposed the now widely accepted argument approach to validation [205], which recognises that no one study can confirm the validity of a measure and each validation study has the potential to add to the certainty of the conclusions we can form as to the validity of the measure. Studies using different comparator measures, in different populations and in different research settings will add weight to the validity portfolio of the ICECAP measures and allow greater understanding of their measurement properties.

9.7.1. Research with a greater spread of comparators

Future validity and responsiveness analyses, both within and outside of a randomised controlled trial, will benefit from the use of a broader range of comparator and anchors measures. Measures that allow a wider nomological network to be defined and a greater breadth of hypotheses to be tested will advance the understanding of the validity and responsiveness of the ICECAP measures. Validation against measures of social connectedness, happiness, enjoyment and independence would enable a better understanding of the associations of the ICECAP measure and greater certainty in the validity conclusions reached.

This should be coupled with the use of both subjective and objective measurement that would also provide increased certainty in the conclusion formed. The majority of the research to

date has used subjective patient reported outcome measures. Many of the constructs assessed by these measures do not have a clear objective assessment option (for example, there is no objective assessment of happiness). However, for those that can be measured objectively, such as physical functioning or bodily strength [321], use of such measures would be of value to future analyses.

9.7.2. Responsiveness and causality

The most urgent area for future research is responsiveness. As discussed earlier the results of this responsiveness analysis provide an indication of how the ICECAP measures respond to changes in health, but provide little indication of how it responds to changes in individuals underlying capability. Further research is required to assess both of these points. Trials with similar outcomes to the trials that we have used here would allow further assessment of the changes in ICECAP scores with the changes in health. For this analysis more objective measures of health, in addition to patient reported health, would be useful in drawing firm conclusions.

The assessment of responsiveness of the measure to changes in capability is more challenging. As described in Chapter 2 a limited number of capability measures, suitable for use in a health context exist. Furthermore, the responsiveness or validity of these measures is either un-researched or under-researched. Therefore, the potential of assessing the responsiveness of the ICECAP measures, by using one of these measures to form anchor groups seems unwise. An option is to assess responsiveness of the measure using a wide variety of anchor measures, which include assessments of factors other than health that determine capability, identified by Grewal[109] and Al-Janabi [115]. Such an analysis would

also include using measures of social-connectedness, physical and financial security, independence and happiness in the formation of anchor groups. If this was to be completed in a trials environment it seems likely that trials where health is not the primary outcome would need to be recruited, possibly including social care or psychological health trials.

A final area for future research is in populations with low baseline capability and populations who experience changes in capability from a low starting point. As described above, this analysis assessed capability at the upper end of the ICECAP range, where changes in capability held less value in the tariff score. Assessment of the responsiveness of the measure in those with low levels of capability would add greater context to these results.

9.7.3. Selection of measures

An emerging theme in the qualitative research was how measures are selected for inclusion in randomised controlled trials. This research allowed the identification of the considerations undertaken when choosing measures, as well as who the main agents are in making this choice. As has been shown in this discussion this knowledge was useful in understanding and contextualising the data pertaining to face validity, but it is also of use independent of a validity analysis. The analysis showed provides future developers of measures with important information on the qualities of measures that are important to the potential users of measures. Also identified is the possibility of a misunderstanding of who holds expertise in the area of quality of life; this may be important for the governance and running of trials.

The assessment of how measures are selected is thought to be the first analysis of its kind and as noted above data saturation was not achieved on this emergent theme. Further qualitative research is required on this topic. Individual interviews would be well complemented by

focus groups where the different agents in the selection process are included. It is likely that the spontaneity that arises from the more challenging and dynamic context of a focus group [244], where participants respond to each other would allow the process of choice in a clinical trial to be better understood. Furthermore, a study which observes the process of measure selection in a trial planning meeting would be of great benefit to understanding the process.

9.8. Conclusion

There is growing interest in the application of Sen's capability approach in health and health economics research [66]. The ICECAP-A [115] and ICECAP-O [100] are two measures that have been developed to assess capability in a health research setting [159,163]. This thesis reports a qualitative assessment of the content validity of the ICECAP-A and quantitative investigations of the validity and responsiveness of both ICECAP measures.

Qualitative findings indicate that health research professionals view the ICECAP-A as measuring content which is relevant to the assessment of capability, or a broader conceptualisation of well-being, and feasible for use in health research. The breadth of the measures was frequently noted, along with the observation that the measure would be of use in addition to, rather than as a replacement for, existing measures of quality of life.

The quantitative results largely confirm *a priori* expectations, providing a strong indication that the measures are valid for use in a health research environment. Moderate correlations and item-by-item associations of the ICECAP measures with measures of physical and psychological health reflect and extend past findings, by providing data from an interventional research setting. The longitudinal data provides the first analysis of the responsiveness of the ICECAP measures. The results indicate that the measures respond to changes in health and highlight the need for further confirmatory research in this area.

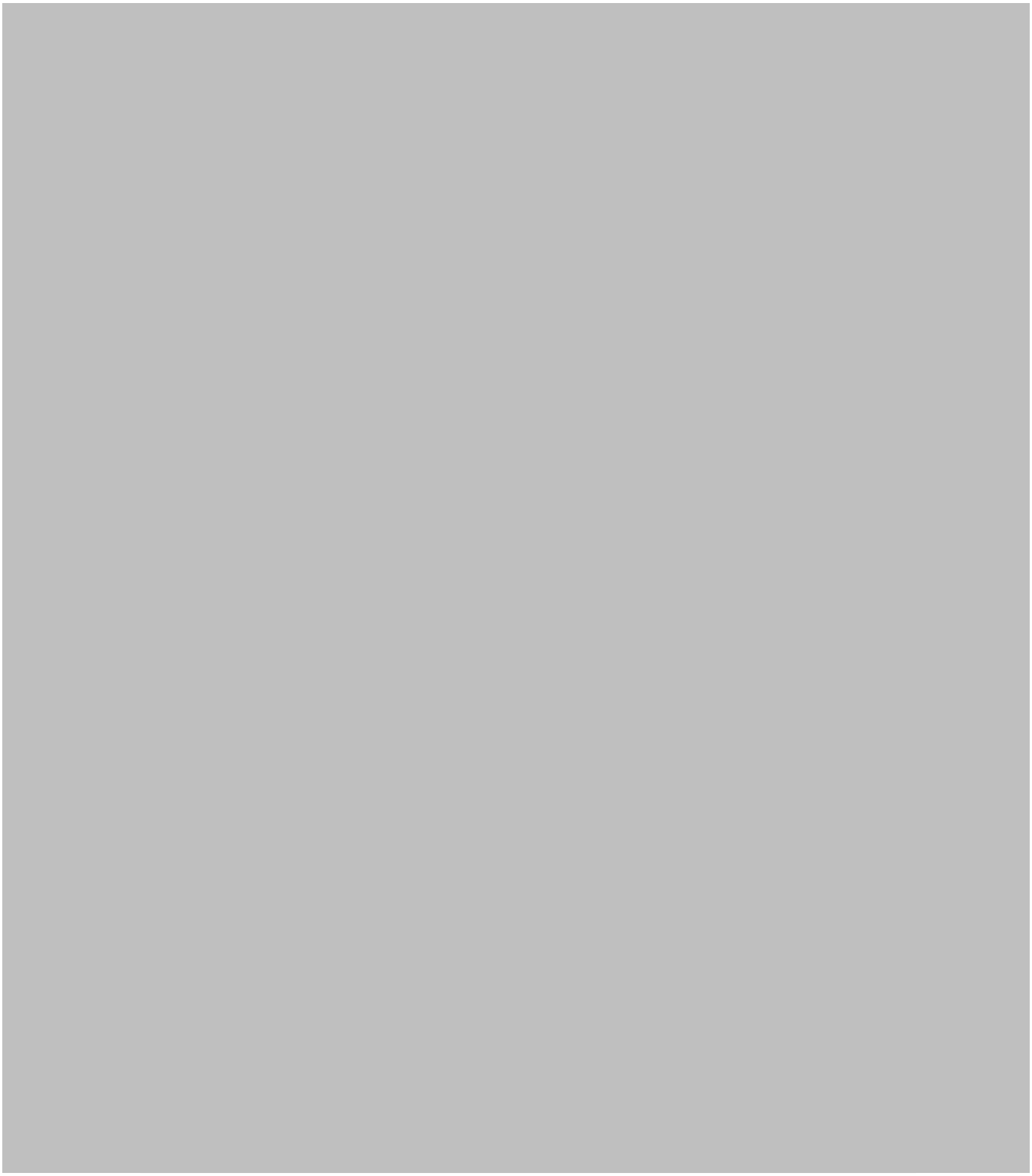
Together, the qualitative and quantitative results provide greater confidence in the conclusions that can be drawn from the ICECAP measures and show that changes that occur in an interventional health research setting will be detected. In doing so, this research has made a

substantial contribution by filling many of the pre-existing gaps in the literature, as well as identifying future directions for research.

APPENDICES

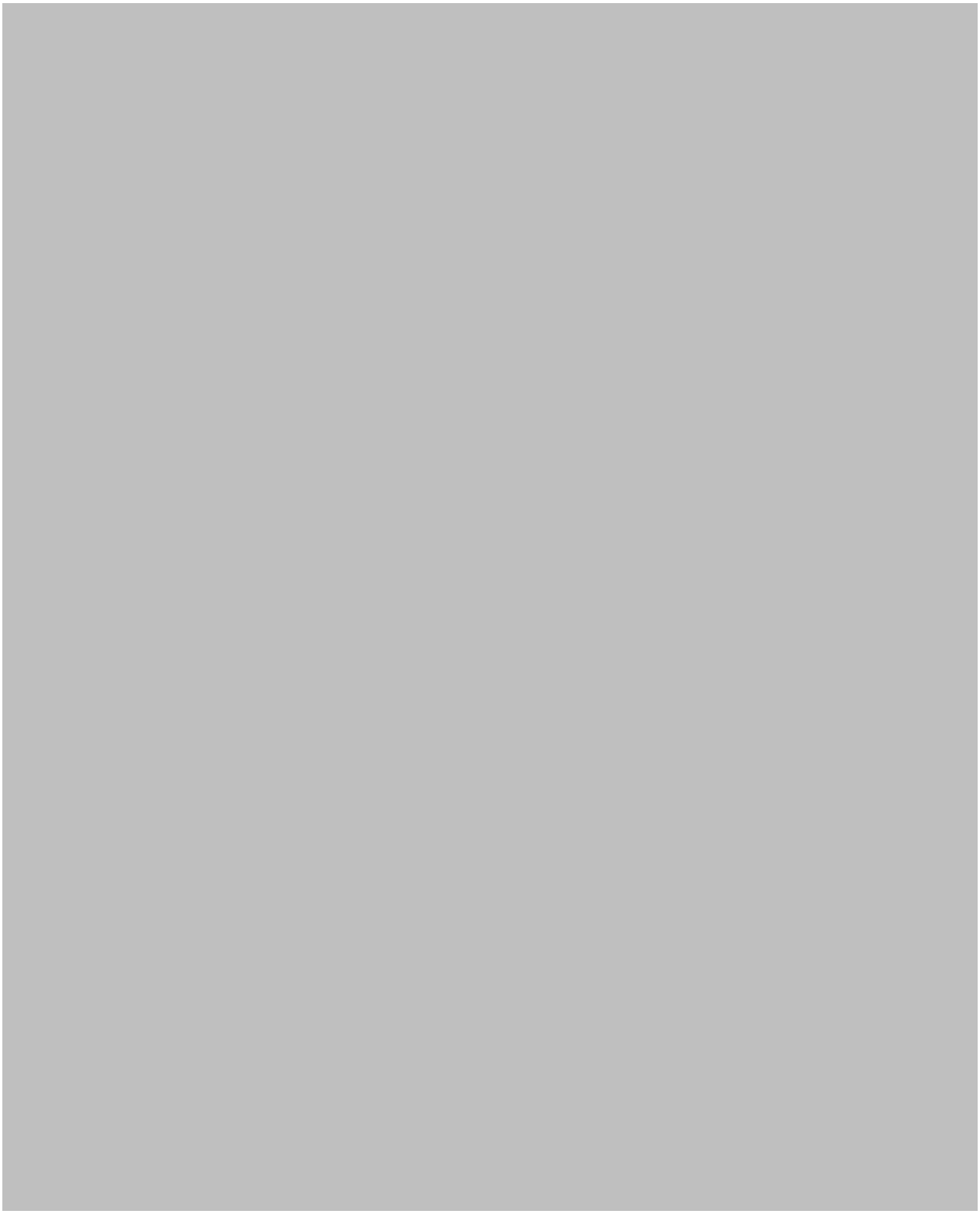
Appendix 1

The ICECAP-O measure



Appendix 2

The ICECAP-A measure



Appendix 3

Text of key person search e-mail

Dear ,

I am a PhD student based at the MRC Midlands Hub for Trials Methodology Research, the University of Birmingham, conducting research into the validity of the ICECAP-O and ICECAP-A capability measures. As part of this research I am currently completing a systematic review of validation studies of the ICECAP measures.

As part of the search strategy I am contacting people who may have knowledge of studies that have attempted to validate the ICECAP measures. Prof. Jo Coast and Dr. Hareth Al-Janabi suggested that I contact you as someone who has either registered use of the ICECAP measure via the University of Birmingham website or attended the ICECAP workshop in 2011.

It would be very helpful if you could notify me of a) any studies that you may have undertook looking at the validity of the ICECAP measures or b) any studies that you are aware of that provide validity evidence on the ICECAP measures (either published or in press). Validity evidence can be both quantitative and qualitative and may be included in a study where the primary aim is not the validation of a measure.

I am extremely grateful for any help you may be able to give.

Yours sincerely

Tom

Appendix 4

Detailed summary of review studies

Author (Year)	Country	Participants: General population or patients	Participants: demographic and health characteristics	Participants: Number in study	Study: Design	Study: sampling strategy	Study: comparators used	Study: analyses used	Study limitations reported by authors of study
Articles providing information on the ICECAP-O									
Coast (2008)	UK	General population	**Mean age**: 74.6 (6.43) **Female**: 56.2% **Mean ICECAP**: not reported **Mean EQ-5D-3L**: 0.76 (0.27)	n=315	Cross-sectional study.	Sample stratified based on postcode.	Socio-demographic. EQ-5D-3L. Disability and Pain. Social support and contact with others.	Chi-squared. One-way ANOVA.	ICECAP administered a couple of months after the comparator measures. Partial assessment of construct validity due to comparators available. No ability to distinguish between cause and effect.
Couzner (2012)	Aus	Patients of an outpatient day rehabilitation unit or a residential transition care unit, with 3 months of hospital admission.	**Mean age**: outpatient rehab 74.87 (7.17), transition care 80.69 (6.27) **Female:** outpatient rehab 45%, transition care 59% **Mean ICECAP**: outpatient rehab 0.82 (0.15), transition care 0.79 (0.16) **Mean ICECAP:** outpatient rehab 0.54 (0.25) 0.49 (0.3), transition care 0.79 (0.16)	n=82	Cross sectional study.	Patients either visiting outpatient or residential in transition care facility were invited to participate.	EQ-5D-3L Care transition measure (CTM-3).	T-test. One way ANOVA. Chi-squared.	Small sample size. UK versions of ICECAP-O and EQ-5D-3L and algorithms were used. The number of comparators in the study limited the scope of analysis.

Author (Year)	Country	Participants: General population or patients	Participants: demographic and health characteristics	Participants : Number in study	Study: Design	Study: sampling strategy	Study: comparators used	Study: analyses used	Study limitations reported by authors of study
Couzner (2013)	Aus	General population and post-acute hospital patients	**Mean age:** not reported **Female**: patients 41%, general population ICECAP 62%, general population EQ-5D-3L-3L 49% **Mean ICECAP**: general population 0.795 (0.17), patients 0.753 (0.18) **Mean EQ-5D-3L:** general population 0.789 (0.02), patients mean 0.595 (0.2)	n=1260	Cross-sectional study.	Patients were recruited from outpatient rehabilitation centre. ICECAP-O from Health Omnibus Survey and EQ-5D-3L from online panel survey.	EQ-5D-3L	T-test ANOVA	Limited socio-demographic data available. Different modes of questionnaire administration may have affected results.
Davis (2012)	Canada	Patients visiting falls prevention clinic	**Mean age**: 79.3 (6.2) **Female**: not reported **Mean ICECAP**: 0.815 (0.177) **Mean EQ-5D-3L**: 0.701 (0.291)	n=215	Cross-sectional study.	All patients visiting clinic invited to participate.	EQ-5D-3L.	Spearman's correlations. Paired Wilcoxon sign-rank test. Exploratory factor analysis.	EQ-5D-3L administered first to all patients. Language of ICECAP may not be appropriate for Canadian participants and UK scoring algorithms were used. Possible influence of completing measures in controlled clinic environment.

Author (Year)	Country	Participants: General population or patients	Participants: demographic and health characteristics	Participants : Number in study	Study: Design	Study: sampling strategy	Study: comparators used	Study: analyses used	Study limitations reported by authors of study
Davis (2012)	Canada	Patients visiting falls prevention clinic	**Mean age**: 79.3 (6.2) **Female**: not reported **Mean ICECAP**: 0.815 (0.177) **Mean EQ-5D-3L**: 0.701 (0.291)	n=215	Cross-sectional study.	All patients visiting clinic invited to participate	EQ-5D-3L Physiological Profile Assessment (PPA). Short Physical Performance Battery. MMSE. Instumental Activities of Daily Living (IADL)	Pearson or Spearman correlations. Stepwise linear regression models.	Cross-sectional so unable to prove causation. Limited sample size (although bootstrapping used). Due to specificity of sampling results may not generalise.
Flynn (2011)	UK	General population	**Mean age**: not reported **Female**: not reported **Mean ICECAP**: 0.832 (0.123) **Mean EQ-5D-3L**:	N=809	Cross sectional study	All 65 year olds and over returning a city wide quality of life survey.	Numerous socio-demographic variables, including whether receiving benefits or caring for a person.	Univariate statistics. Multivariable regression.	
Horwood (2013)	UK	Pre and post-operative patients with osteoarthritis of hip or knee	**Mean age:** 70 **Female**: 70% **Mean ICECAP:** pre-op patients 0.77; post-op patients 0.82. **Mean EQ-5D-3L:**	N=20	Qualitative think aloud study.	Purposive sampling	n/a	A thematic analysis and an item-by-item analysis of response problems.	Think aloud methodology is dependent on the ability of the informants to verbalise their thoughts. The clinical population used (although complements general populations used in development.

Author (Year)	Country	Participants: General population or patients	Participants: demographic and health characteristics	Participants: Number in study	Study: Design	Study: sampling strategy	Study: comparators used	Study: analyses used	Study limitations reported by authors of study
Makai (2012)	Holland	Psycho-geriatric patients	**Mean age**: nursing version 82 (9.1), family version 82 (7.3) **Female**: nursing version 68%, family version 67% **Mean ICECAP**: nursing version 0.5 (0.2), family version 0.43 (0.17) **Mean EQ-5D-3L**: nursing version 0.49 (0.21), family version 0.46 (0.2)	n=122	Cross-sectional study.	Convenience sample of those in restraints and a random sample of non-restrained controls.	EQ-5D-3L. Cantril's ladder. HADS. Care dependency scale (CDS). Overall life satisfaction.	Chi-sqaured. Mann-Whitney U. OLS regression.	Small sample size. Limited analysis due to limited comparators available. Possible methodological issues with multiple imputation used.
Makai (2013)	Holland	Patients 3 months post-admission to hospital	**Mean age:** 76.21 (6.79) **Female:** 53.8% **Mean ICECAP:** 0.84 (0.14) **Mean EQ-5D-3L:** 0.8 (0.17)	n-275	Cross sectional study	All patients admitted to a hospital in a 5 month period were approached to participate	EQ-5D-3L. Cantril's ladder. Geriatric depression scale (GDS-15). SF-20	Correlation analysis. T-test. One-way ANOVA. Stepwise multi-variate regression.	Sample not representative of elderly post-admission populations (more frail).
Ratcliffe (2013)	Aus	General population	**Mean age:** not reported **Female:** carers 77%, non-carers 61% **Mean ICECAP:** carers 0.848 (0.123), non-carers 0.838 (0.147) **Mean EQ-5D-3L:**	n=786	Cross-sectional study	Health Omnibus Survey	Numerous socio-demographic variables including whether they are informal carers.	Kruskal-Waliis analysis of variance. Mann-whitney U test.	Study nested with larger study and framing influences are possible. No assessment of hours per week caring. Small sample size in the carers group. UK version of ICECAP and scoring algorithm used.

Author (Year)	Country	Participants: General population or patients	Participants: demographic and health characteristics	Participants: Number in study	Study: Design	Study: sampling strategy	Study: comparators used	Study: analyses used	Study limitations reported by authors of study
Ratcliffe (2011)	Aus	Patients of an outpatient day rehabilitation unit, inpatient medical rehabilitation or a residential transition care unit.	**Mean age**: 75.8 (IQR 69-84) **Female**: 59.7% **Mean ICECAP**: 0.779 (0.154) **Mean EQ-5D-3L**: 0.526 (0.297)	n=181	Cross sectional study.	All eligible patients attending one of the three treatment centres were invites to participate.	EQ-5D-3L. CTM-3. Herth Hope Index. Modified Rankin Scale.	Chi-squared. Spearman's correlation.	Small sample size. A small number of participants (n=22) were under 65 when the ICECAP-O was administered to them. UK versions of ICECAP-O and EQ-5D-3L and algorithms were used.
Articles proving information on the ICECAP-A									
Al-Janabi (2012)	UK	General population	**Mean age**: 51.7 (18.2) **Female:** 62% **Mean ICECAP:** 0.832 (0.157)* **Mean EQ-5D-3L:** 0.815 (0.245)*	n=418	Cross-sectional study	Sampled using a two stage stratified sampling procedure.	Numerous socio-demographic variables, including life events. EQ-5D-3L and other health variables	Chi-squared. One-way ANOVA.	Slight over sampling of female participants.
Al-Janabi (2013)	UK	General population	**Mean age:** not reported **Female:** 53% **Mean ICECAP:** not reported **Mean EQ-5D-3L:** not reported	n=34	Qualitative think aloud study.	Purposive sampling	EQ-5D-3L	Qualitative constant comparative analysis	Potential queries around the effect of think aloud methodology. Very few informants were in bad health states.

Appendix 5
Invitation e-mail

Dear <Name>

I am a PhD student at the University of Birmingham, conducting research on Quality of Life patient reported outcome measures. A part of my research I am carrying out a number of short interviews with clinical trialists, medical clinicians involved in research, research nurses and other "frontline" researchers.

<Name of individual> recommended you as an excellent person to speak to and I obtained your contact details through <insert source>. I would be very interested in speaking to you as <role of potential participant> who has involvement in research. Your professional experience could be very informative to this project.

From these interviews I am looking to gain a greater understanding of quality of life measures in a research setting, how quality of life data is handled and specifically I am looking for opinions on a number of quality of life measures currently in use.

If you agree to take part the interview will last approximately 40 minutes, and no longer than an hour and can be conducted at a location most convenient for you (most likely your place of work) and a time most convenient for you. Attached is an information sheet giving you more information about the research.

The research is confidential. Once the interview is recorded a unique identity code will be allocated to the information you have provided, allowing us to keep your details and the recording separate. Furthermore, we secure the information you provide by storing it on a secured computer network which only the lead researcher has access to. Data will be stored for 10 years. Data, in this form, will be analysed in collaboration with colleagues at Monash University, Australia see information sheet for more details.

I have a contact number for you (<number>) and I will call you within the next week to discuss this further. Alternatively, you could contact me on <number>. If you would rather not be contacted please e-mail me at tjk962@bham.ac.uk.

Yours sincerely

Thomas Keeley,
Doctoral Researcher
Health Economics Unit / MRC Midlands Hub for Trials Methodology Research
The University of Birmingham

Appendix 6

Informant information sheet

UNIVERSITY OF BIRMINGHAM | MRC Hubs for Trial Methodology Research Midlands Hub

Quality of Life measures study

Participant information sheet

We would like to invite you to take part in our research study. Before you decide it is important to understand why the research is being done and what it would involve for you. We suggest that you take a couple of minutes to read through the information below and ask any questions about the research that you may have.

Brief summary of research

Quality of life is becoming an increasingly important outcome. Short quality of life questionnaires are now frequently used in almost all clinical and public health trials. At the University of Birmingham we are conducting research that will allow us to understand what is required from quality of life measures used in research and what the opinions of research professionals are in regards to quality of life assessment.

This research is being carried out by **the University of Birmingham** and is funded through **the UK Medical Research Council. The study has been approved by the Science, Technology, Engineering and Mathematics Ethical Review Committee of the University of Birmingham.**

Questions you may have

Why have I been invited to take part?

You have been invited to participate in this project because of your professional role as someone who works within a health research setting.

Do I have to take part?

No, you are under no obligation to take part. Participation is entirely voluntary; however, this research is dependent on the goodwill and cooperation of those who take part. Participation in this research does not mean you will be required or obliged to participate in any future research.

What is the research about?

We are asking you to take part in a research interview. We are interested in three main areas: 1) your opinion of the use of quality of life assessment in health research, 2) your

opinions on two specific quality of life measures and 3) how your study uses quality of life data once collected.

What will the information be used for?

The information will allow us to investigate the opinions of research professionals towards quality of life measurement and the factors affecting choice of instrument. Specifically the results will be used to inform the development and testing of the quality of life measures discussed in the interview.

This research will form a major part of a PhD thesis. Results will also be disseminated through seminars, academic journals and conferences, adding to a growing body of research of quality of life measures in health research.

What will happen if I take part?

If you agree to take part the interview will take roughly 45 minutes, and no longer than one hour. During this time you will be asked to discuss a series of topics (described above) related to quality of life and quality of life measures. At no point will you be asked to discuss your own quality of life; however you may find it useful to use personal examples. This is not a test as it is your opinions which we are interested in.

Am I free to pull out of the research?

You are free to withdraw from the interview and the research, without the need to provide a reason. Upon withdrawal you have the option of asking for the data you have given to be destroyed. If you withdraw during the interview we will ask you then whether you want to withdraw from the research as a whole and have your data to be destroyed. If you decide to withdraw from the research after the interview, please ensure to notify us within a week of the interview; the data you have provided will then be destroyed. If you notify us after this time period it is possible the data may already have been analysed and used to inform future interviews.

Is the research confidential?

Yes, the research is confidential and will follow ethical and legal best practice. Once the interview is recorded a unique identity code will be allocated to the information you have provided. This allows us to keep your details and the recording separate.

Furthermore, we secure the information you provide by storing it on a secured computer network which only the lead researcher has access to. Data will be stored at the University of Birmingham for 10 years.

In any publications or writing resulting from this research neither your name nor the name of the organisation that you work for will be used.

The data collected will be analysed in collaboration with colleagues at Monash University, Australia. Researchers within this institution have expertise on quality of life measures research that will be informative to the project.

Where can I get more information about the study?

What is the complaints procedure?

Sources of support

If you feel you need to talk to someone about any of the topics raised in this interview we suggest you contact the National Counselling Directory at http://www.counselling-directory.org.uk

What now?

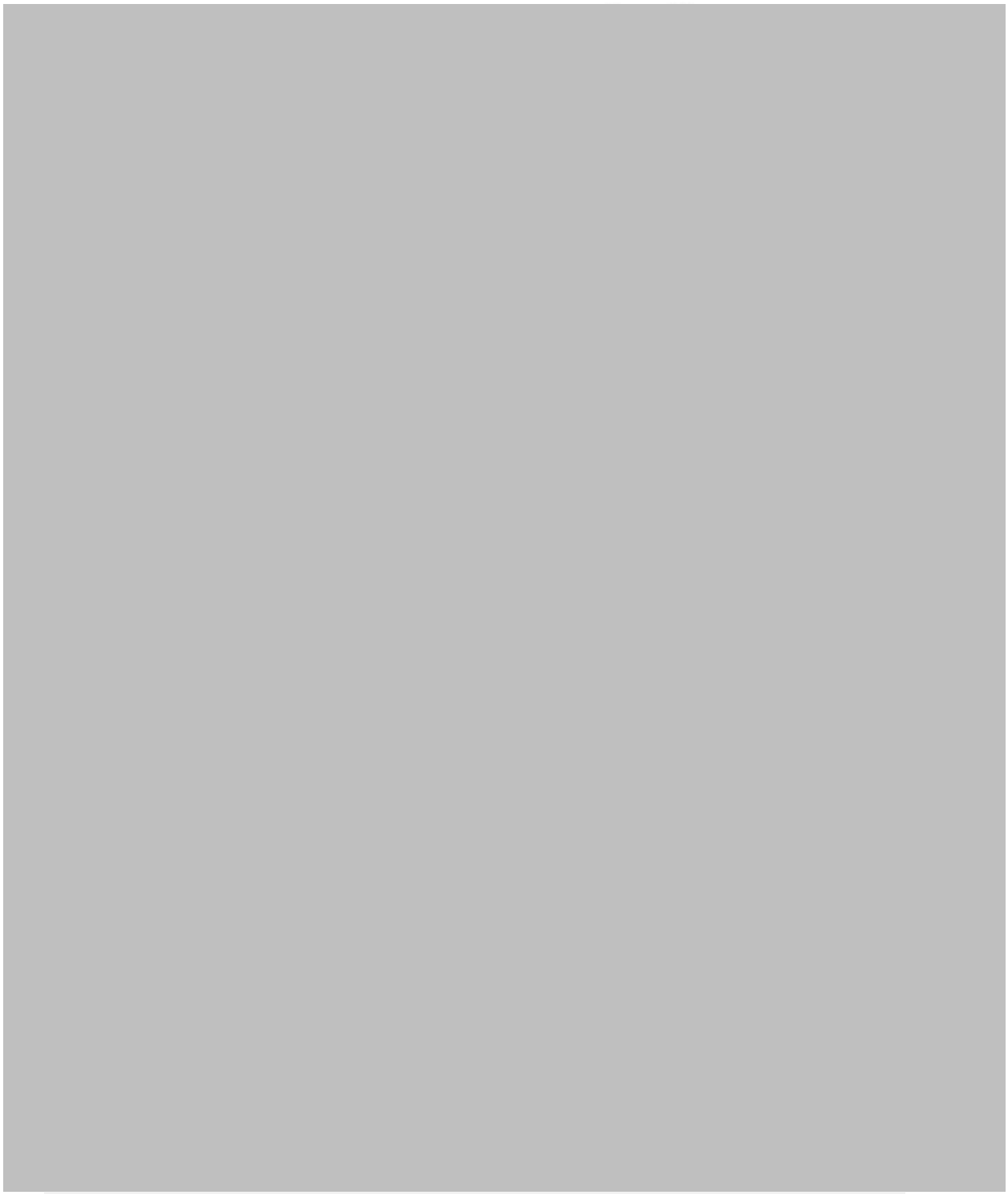
You are free to take your time to consider whether you would like to participate.

If you are happy to participate please sign the consent form.

Appendix 7

Ethical approval for qualitative work

UNIVERSITY OF BIRMINGHAM



Appendix 8

Qualitative topic guide

Topic Guide (Version 3)

Study recap:

"The aim of this study is to gain a greater understanding of quality of life measures in a research setting, how quality of life data is handled in research and I am also looking for your opinion on a couple of quality-of-life measures currently in use. There are no right or wrong answers. The interview will last roughly an hour and you are free to stop the interview at any point without providing a reason. And, just to reiterate, all the research is completely confidential and we take steps to secure the data."

Background information:

Can I start by asking what your role is here? Can you tell me a little bit about the research you are involved in? Could you tell me a little bit about that? Have you had experience of using or developing a quality-of-life measure?

Quality of Life:

What does the term Quality of Life mean to you? What things might affect someone's quality of life? What impact does someone's health have on their quality of life? Can people with poor health have lives of good quality? **How does the diseases and illnesses that you work with impact upon patients' quality of life?**

Is quality-of-life important to the research that you are involved in? How is quality of life measured in your trials? Do you feel that the things that affect quality of life, which we have talked about, are captured in the measures used? If not which? Who decides which measure is to be used?

Amplify: Tell me a little more about that…. Are you able to give me an example.

Explore: In what way does that effect….. To what extent does…

Explanation: Why do you think that is…

Clarification: What do you mean when you say "……"

Opinions of EQ-5D and ICECAP-A measures:

"In this part of the interview I am looking for your opinions on two Quality of Life measures. So, could I ask you to look at this measure, I'll just give you a couple of moments to read through it."

Layout?
Length?
General wording?
Wording of response?
Anything standout?

Now considering the questionnaire as a whole, **What are your first impressions of this measure?** How do you think patients would find completing this measure? How long do you think it would take to complete? Are there any drawbacks that you can see? How do you think it compares to other measures that you know of? **Considering all the questions together – do you feel that this measure offers coverage of the aspects of quality of life that we were discussing earlier?**

Now I would like you to focus on the individual questions. **Are there any questions that you feel to be inappropriate? Follow-up.** Are there any questions which you think patients would find more difficult than others to answer? Do the topics of any of the questions stand out to you?

Would this measure be of use in the research that you are involved in? Why is that? How receptive do you think your colleagues would be to using this measure?

Placing both in front of participant **Now that you have seen both, how do the two measures compare? Would you be more inclined to use one over the other?** What would be required to convince you to us this measure? Do you think there are areas or times when one of these measures may be more useful or appropriate?

Handling of QoL data in trials

Now thinking about how Quality of Life is handled in trials that you have worked with/ or do currently work with. **Once a quality of life questionnaire is completed, what happens to it then?** Nurse: How frequently do you look at QoL forms that patients have completed? Do you act on what the forms tell you? Clinicians: In what circumstances would you want to know what is on the QoL questionnaire for a patient of yours? Trialists: Do you think that research nurses should be looking at QoL forms? Do you think they should act upon it?

Appendix 9

Qualitative analysis codes

1) Professional details

- **1.1) Current role**
  - 1.1.1) Research area
  - 1.1.2) Role
- **1.2) Background/experience**
- **1.3) Informant & QoL measurement**
  - 1.3.1) QoL experience
    - 1.3.1.1) Measured used
    - 1.3.1.2) Knowledge
  - 1.3.2) Involvement in measure selection

2) QoL

- **2.1) Beliefs**
  - 2.1.1) Physical
    - 2.1.1.1) Pain
    - 2.1.1.2) Functioning
      - 2.1.1.2.1) Mobility
        - 2.1.1.2.1.1) Mobility independence
    - 2.1.1.3) Health
      - 2.1.1.3.1) Treatment
        - 2.1.1.3.1.1) Side-effects
  - 2.1.2) Psychological
    - 2.1.2.1) Psychological outlook
    - 2.1.2.2) Happiness
  - 2.1.3) Social
    - 2.1.3.1) Freedom
      - 2.1.3.1.1) Independence
    - 2.1.3.2) Work
  - 2.1.4) Family and friends
  - 2.1.5) Life
    - 2.1.5.1) Normal life
    - 2.1.5.2) Role achievement
    - 2.1.5.3) Goal attainment
    - 2.1.5.4) Life Events
    - 2.1.5.5) Financial
  - 2.1.6) Broad construct
  - 2.1.7) Capability
  - 2.1.8) Adaptation
- **2.2) Measurement**
  - 2.2.1) Assessment
    - 2.2.1.1) Other assessment methods
    - 2.2.1.2) PROMs

2.2.1.2.1) Who selects
2.2.1.2.1.1) Precedent
2.2.1.2.2) Measures used
2.2.1.2.2.1) Opinions
2.2.1.2.3) Patient disruption
2.2.2) Subjectivity
2.2.2.1) Adaptation
2.2.2.2) Interpretation
2.2.2.3) Completion
2.2.3) Requirement

2.3) Psychometrics

2.3.1) Sensitivity to change
2.3.2) Validity
2.3.3) Reliability
2.3.4) To use
2.3.4.1) Precedent
2.3.5) Subjectivity
2.3.5.1) Adaptation
2.3.5.2) Interpretation

3) Measure assessment

3.1) ICECAP-A

3.1.1) First reaction
3.1.2) Basics
3.1.2.1) Length
3.1.2.2) Layout
3.1.2.3) Understandable
3.1.2.4) Patient friendly
3.1.2.4.1) Ease of completion
3.1.2.4.1.1) missing data
3.1.2.4.2) Patient focused
3.1.2.4) Of use
3.1.2.4.1) Of use colleagues
3.1.2.5) Appropriateness
3.1.3) Wording
3.1.3.1) Wording capability
3.1.4) Focus
3.1.4.1) Psychological vs Physical
3.1.4.1.1) Psychological
3.1.4.1.2) Functioning
3.1.4.2) Timeframe
3.1.4.3) Breadth
3.1.4.3.1) Questionnaire breadth
3.1.4.3.2) Item breadth
3.1.4.4) Relevance

3.1.4.5) Question duality
3.1.4.5) Coverage
3.1.5) Response options
3.1.5.1) Number
3.1.5.2) Unachievable
3.1.6) Attributes/Constructs
3.1.6.1) Stability
3.1.6.2) Attachment
3.1.6.3) Autonomy
3.1.6.4) Achievement
3.1.6.5) Enjoyment and pleasure

3.2) EQ-5D

3.2.1) First reaction
3.2.2) Basics
3.2.2.1) Length
3.2.2.2) Layout
3.2.2.3) Understandable
3.2.2.4) Patient friendly
3.2.2.4.1) Ease of completion
3.2.2.4.1.1) missing data
3.2.2.4.1) Patient focused
3.2.2.4) Of use
3.2.2.4.1) Colleagues
3.2.2.5) Appropriateness
3.2.3) Wording
3.2.4) Focus
3.2.4.1) Psychological vs Physical
3.2.4.1.1) Psychological
3.2.4.1.2) Functioning
3.2.4.2) Timeframe
3.2.4.3) Breadth
3.2.4.3.1) Questionnaire breadth
3.2.4.3.2) Item breadth
3.2.4.4) Relevance
3.2.4.5) Question duality
3.2.4.6) Coverage
3.2.5) Response options
3.2.5.1) Number
3.2.5.2) Unachievable
3.2.6) Attributes/Constructs
3.2.6.1) Mobility
3.2.6.2) Sefl-Care
3.2.6.3) Usual Activities
3.2.6.4) Pain/Discomfort
3.2.6.5) Anxiety/Depression

Appendix 10

OPEN ACCESS Freely available online

PLOS | ONE

A Qualitative Assessment of the Content Validity of the ICECAP-A and EQ-5D-5L and Their Appropriateness for Use in Health Research

Thomas Keeley[1,2*], Hareth Al-Janabi[2], Paula Lorgelly[3], Joanna Coast[2]

1 MRC Midlands Hub for Trials Methodology Research, Health and Population Sciences, University of Birmingham, Birmingham, United Kingdom, **2** Health Economics Unit, Health and Population Science, University of Birmingham, Birmingham, United Kingdom, **3** Centre for Health Economics, Monash University, Victoria, Australia

Abstract

Purpose: The ICECAP-A and EQ-5D-5L are two index measures appropriate for use in health research. Assessment of content validity allows understanding of whether a measure captures the most relevant and important aspects of a concept. This paper reports a qualitative assessment of the content validity and appropriateness for use of the eq-5D-5L and ICECAP-A measures, using novel methodology.

Methods: In-depth semi-structured interviews were conducted with research professionals in the UK and Australia. Informants were purposively sampled based on their professional role. Data were analysed in an iterative, thematic and constant comparative manner. A two stage investigation - *the comparative direct approach* - was developed to address the methodological challenges of the content validity research and allow rigorous assessment.

Results: Informants viewed the ICECAP-A as an assessment of the broader determinants of quality of life, but lacking in assessment of health-related determinants. The eq-5D-5L was viewed as offering good coverage of health determinants, but as lacking in assessment of these broader determinants. Informants held some concerns about the content or wording of the Self-care, Pain/Discomfort and Anxiety/Depression items (EQ-5D-5L) and the Enjoyment, Achievement and attachment items (ICECAP-A).

Conclusion: Using rigorous qualitative methodology the results suggest that the ICECAP-A and EQ-5D-5L hold acceptable levels of content validity and are appropriate for use in health research. This work adds expert opinion to the emerging body of research using patients and public to validate these measures.

Citation: Keeley T, Al-Janabi H, Lorgelly P, Coast J (2013) A Qualitative Assessment of the Content Validity of the ICECAP-A and EQ-5D-5L and Their Appropriateness for Use in Health Research. PLoS ONE 8(12): e85287. doi:10.1371/journal.pone.0085287

Editor: Ulrich Thiem, Marienhospital Herne - University of Bochum, Germany

Received June 13, 2013; **Accepted** December 4, 2013; **Published** December 19, 2013

Funding: The work is funded through the Medical Research Council Midlands Hub for Trials Methodology Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: TJK962@bham.ac.uk

Introduction

The ICECAP-A is a relatively new index measure. Its theoretical grounding is in Amartya Sen's work on functioning and capability [1–4] which encourages a broad evaluative space through a focus on what a person is able to do and who they are able to be, rather than what they actually do and who they become [5]. The descriptive system of the ICECAP-A capability measure, formed through in depth interviews with the general public, defines quality of life as consisting of: Stability, Attachment, Autonomy, Achievement and Enjoyment [6]. The measure assesses capability by asking whether a person "can" or is "able" to achieve particular states. The ICECAP-A measure was developed in the UK and is already being used within health research in a number of countries including Australia and Canada.

The EQ-5D-5L is a generic preference based outcome measure, which measures health-related quality of life. The descriptive system remains the same as the original EQ-5D-3L measure: mobility, self-care, usual activities, pain and discomfort and anxiety and depression [7], with the number of response options in each dimension increased from three to five. This aims to increase the responsiveness of the measure and reduce ceiling effects [8–12]. The EQ-5D-5L has been translated into 97 languages and work to elicit value sets has begun in a number of countries [13].

Content validity is the extent to which a descriptive system of a measure "represents the most relevant and important

aspects of a concept in the context of a given measurement application" [14]$_{p.743}$. Assessment of the content coverage of a measure allows understanding about inferences that can be drawn from the results of a measure. Here, content validity concerns whether the ICECAP-A and the EQ-5D-5L assess the most important and relevant attributes of quality of life and health-related quality of life, respectively.

When measuring a non-tangible concept, the questions of what is relevant and what should be measured is a longstanding methodological challenge in content validation [15,16]. This challenge is particularly important when attempting to validate the content of quality of life and health-related quality of life measures, where no universally accepted definition exists. An analysis that seeks to answer the question "does the measure sample the important and relevant dimensions of a construct?" requires clarity regarding these important and relevant dimensions. Therefore, qualitative analyses of the content validity of a measure should seek to assess not only the opinions of the measure under consideration, but also the informant's conceptualisation of the concept the measure seeks to assess. Having an understanding of the informant's conceptualisation of the concept creates the potential for a fuller understanding of the opinions of the measure and a firmer conclusion as to the content validity of the measure.

Rigorous and transparent qualitative methodology, absent from much of the research associated with quality of life measure validation, provides a suitable way for assessing content validity [17–19]. Qualitative content validation can be completed with the public, patients and experts in relevant fields acting as informants. Patients have first hand, personal experience of both the concept and how it might be affected by different situations. They are in a position to provide an insider's – emic – perspective. Experts, such as clinicians or researchers can provide an outsider – etic – perspective. They have the advantage of observing both a number of individuals in different situations and how the construct of interest manifests itself in different individuals [14]. While there is an emerging quantitative evidence base of validity amongst patients and public for both the EQ-5D-5L [8,20,21] and ICECAP-A [22], qualitative work using experts as informants is needed to allow triangulation of data from different perspectives and different methodologies [14,19].

This paper reports a qualitative assessment of the content validity of the EQ-5D-5L and ICECAP-A measures and their appropriateness for use in health research.

Methods

Informant selection

Informants were recruited from the UK and Australia to provide an international assessment of these measures. Using maximum variation sampling [23] informants were purposively selected from: 1) clinical and public health trial experts ("trialists"); 2) medical doctors involved in research; 3) researchers with regular participant contact ("frontline researchers"); and 4) health economists working within a trial setting. These groups were selected to sample across a broad spectrum of professional experiences and research perspectives.

Invitation emails stating the aim of the research, the potential burden upon the informant and the ethical approval gained for the study were sent to potential informants. Snowball recruitment, whereby previously interviewed informants were asked to give recommendations of other potential informants, was used to recruit three frontline researchers. Recruitment was stopped when data saturation was identified.

Ethics statement

The study protocol, which included participants providing written consent prior to the interview, was approved by the University of Birmingham Science, Technology, Engineering and Mathematics Ethical Review Committee (ERN_11-0575).

The interview and analysis process

Interviews were broadly partitioned into two parts using a semi-structured topic guide designed to facilitate breadth and depth of discussion. A two stage investigation of content validity, termed the comparative direct approach, was developed based on this partitioning. The development of this approach sought to address the methodological challenge discussed above by providing a useful structure within which a thematic analysis, grounded in the data could be completed.

The first part of the interview assessed informants' understanding of quality of life as a concept, what influenced this understanding and how the diseases they worked with professionally affected their perspective. Content mapping and mining questions were used to encourage breadth and depth [24]. Differences between quality of life and health-related quality of life were explored. When analysing the data from this first part of the interview the informants' descriptions of quality of life were *compared* by the researcher with the descriptive systems of the ICECAP-A and EQ-5D-5L. This comparison enabled identification of the parts of a measure's descriptive system that the informants felt were important.

In the second part of the interview, informants were presented with copies of ICECAP-A and EQ-5D-5L, one after the other, in random order. Informants were encouraged to *directly* discuss the measure's content coverage and its appropriateness for use in their research area. Informants were asked to think back to how *they* defined quality of life and health-related quality of life in the first part of the interview, and assess whether they felt the measure covered *their* conceptualisation. Data from the second part of the interview was analysed to assess informants' opinions about how well the measures covered their own conceptualisation of quality of life. Using the informant's conceptualisation of quality of life as a reference point facilitated the analysis: knowing what the informants understood quality of life to be and what dimensions they held to be important and relevant, allowed greater understanding of their opinions of the measures.

All interviews were conducted by TK, who was not involved in the development of either of the measures under consideration. Interviews were recorded, transcribed verbatim and transcripts were coded using a hierarchical and flexible coding structure. The first version of this coding structure was

formed during the completion of the first analysis batch and was therefore grounded in the data. This coding structure allowed data to be coded under broader themes as well as under more focused categories referring to a specific topic. An iterative, constant comparative, thematic analysis of the transcripts was completed [17,25,26]. Transcripts were analysed in four successive batches. This analysis allowed descriptive and explanatory accounts to be formed and comparisons to be drawn between informants. The iterative nature of the analysis allowed themes which were identified in earlier batches to be analysed and developed in later batches of the analysis. Themes in the data were identified by the authors and developed through the use of the flexible coding structure. TK led the coding and analysis of the data and work was checked regularly by HAJ, PL and JC for consistency and accuracy. The qualitative data analysis computer package ATLAS.ti was used to assist this analysis. Verbatim quotes from informants have been selected to illustrate how informants' accounts were linked to emerging themes. Ellipses (…) were used to denote missing speech. 'Umm', 'err' and repetitions of words, which do not add meaning, were removed without the use of ellipsis.

Findings

In this section the informants' conceptualisations of quality of life and its determinants are presented. The overall perceptions of the measures, followed by an item by item breakdown of the results of the researcher led *comparative* analysis and the informant led *direct* analysis, for the ICECAP-A and EQ-5D-5L are presented in turn.

Interviews

Interviews, lasting between 45 and 90 minutes, were conducted with 17 informants in the UK and Australia between February and September 2012 (see Table 1 for informant characteristics). Interviews were conducted at the informants' place of work. None of the informants were involved in the development of either the ICECAP-A or EQ-5D-5L measures, nor did they hold any professional relationship with TK. Data saturation was identified at interview 14. Three additional interviews were conducted to check saturation and ensure adequate numbers were sampled from each professional role.

Informant conceptualisation of quality of life and health-related quality of life

Physical health was identified by an overwhelming majority of informants as an important determinant of both an individual's health-related quality of life and quality of life. Pain was identified as a particularly pervasive determinant and for informants who worked in cancer research the side-effects of treatment were a particular concern. Psychological health was also seen as having a notable impact on quality of life. Many informants discussed how psychological problems, such as depression, can stem from a physical condition.

I still see health as important…I think when someone has got ill health...it is quite a big determinant. [Frontline researcher, Australia]

Table 1. Informant characteristics.

	Number interviewed (n=17)
Sex	
Male	7
Female	10
Location	
Australia	8
UK	9
QoL measure experience of use	
EQ-5D	15
ICECAP	5
Professional role	
Frontline researcher	4
Trialist	6
Health economist	3
Research doctor	4

doi: 10.1371/journal.pone.0085287.t001

Informant discussions indicated they differentiated between quality of life and health-related quality of life, with the later closely related to physical and psychological health. Quality of life was viewed as a broader construct and the terms "big picture", "multi-dimensional" and "broad" were frequently used. Although physical and psychological health were considered major determinants, informants recognised that people can have adequate or even high quality lives, despite being in poor health states.

…all you see is ill health and states that you don't want to get into, but there are people that get into those states and have a fantastic time. [Trialist, UK]

Relationships with friends and family were viewed as important due to the enjoyment and support they can provide. Informants often described the importance of friends and family from the perspective of losing loved ones. The ability of an individual to lead their normal life was discussed by a sizable minority of informants. Informants attached importance to individuals being able to fulfil the roles within society which they value.

…that they are able to perform social roles that they would normally perform [is important]. [Trialist, Australia]

ICECAP-A

Informants viewed the ICECAP-A as a broad assessment of quality of life, appropriate for use in the research fields in which they worked. It was viewed as a short, uncomplicated measure, suitable for a busy research environment:

It is a lovely length...because…you don't have the time to spend with a long questionnaire. [Frontline researcher, UK]

Most informants felt that the ICECAP-A captured the important determinants of quality of life as described in part one of the interview. The notable exception was that informants felt that it did not directly assess health, which informants had identified as an important determinant.

It is just how far away from health it gets I suppose...I think it is just the distance from health. [Health economist, UK]

Yeh I guess this one is more general...and focuses mostly on the emotional. [Trialist, Australia]

Most informants felt the measure was patient-focused, while a very small minority, who were more likely to be trialists than frontline researchers or doctors, felt the subject matter was too sensitive for patients. There was a consensus that the measure would be favoured in addition to, rather than as a replacement for, existing health-related quality of life measures. Informants felt that a measure that maintained a focus on health-related quality of life was also required.

I wouldn't see it as a replacement for an EQ-5D, but it would certainly complement an EQ-5D type instrument. [Health economist, UK]

A small number of informants showed a level of cognitive struggle in understanding the capability wording of the ICECAP-A measure. For these informants there was some concern about whether the wording would be understood by participants in the studies.

I don't like the "I am able" or "I can", I don't know, it feels as if in some way you are the person with the control , so I CAN have a lot if I want to I can have a lot of love and friendship. [Trialist, UK]

Stability. Prior to viewing the measure in the first part of the interview, informants identified stability in life as an important determinant of quality of life. Living with fear and uncertainty due to a physical condition or illness and the concern that unemployment due to illness can cause, was identified by a number of informants.

You get frightened of taking your medicine. You get frightened of going to sleep, in case you don't wake up. [Research doctor, Australia]

Upon seeing the measure, there was broad acceptance that the Stability item was relevant to the assessment of quality of life and would be influenced by both health and non-health factors. Some informants recognised that the item was assessing a construct that they had previously identified as important.

...it makes sense because...the patients I see are very palliative and they don't have a lot of time. But you can still be settled and secure with months to live. [Frontline researcher, UK]

Attachment. Prior to viewing the measure, informants identified the ability to function in a social context as an important consideration both for the enjoyment and support it provides. The significance to people suffering from illnesses of achieving social contact and the limiting effect that illnesses can have upon ability to achieve social contact was discussed at length.

And in the last year of his life, he died by the cancer, he said...this has been the best year of my life, because until this moment I never realised how loved I've been. [Trialist, UK]

Upon considering the measure concern was raised by a small number of informants about the perceived sensitivity of the subject, while the majority recognised Attachment as being both relevant and seldom assessed.

Well things like love, friendship and support. It is all that thing around social connectiveness and support and intimacy. We as a research group are very interested in that in people with HIV [Research doctor, Australia]

Autonomy. A small number of informants discussed independence as a dimension of quality of life prior to seeing the measure. Informants focused on the ability to do day-to-day activities that were often closely linked with mobility.

...they can't get down to the shops to do their shopping…It is hard, they have got to think about is it feasible to do something that they want to, based on how mobile they are. [Trialist, Australia]

In comparison to the limited discussion prior to viewing the measure, most informants identified the Autonomy item as being of central importance to the assessment of quality of life. There was a consistent view that it was particularly important to elderly people.

...especially with older people that independence is hugely important to them, and that's one of the depressing things for them when they lose that independence I think. [Frontline researcher, UK]

Achievement. The influence upon quality of life of being able to achieve and attain personal goals was not discussed by many informants prior to viewing the ICECAP-A. However, gaining a sense of achievement through work and being able to look back at life with a sense of achievement were discussed briefly by a small number of informants.

...i think he [young cancer sufferer] has kind of condensed it all to "Yeah, I am 25 and I have achieved everything I want"...and he is perfectly sane in what he is saying. [Frontline researcher, UK]

On seeing the measure, Achievement was thought to be more relevant to younger rather than older people. The use of the word "progress" in the item was questioned. For some it focused on the area of paid employment, while those who worked in cancer noted that oncology patients could misunderstand the question as assessing the progress of their cancer. The item was considered by a number of informants as being too broad and some questioned whether the top item was really achievable.

I don't think that I can achieve and progress in all aspects of my life, I would love to be able to. BUT. [Trialist, UK]

Enjoyment. Enjoyment was discussed as an important influence on quality of life from the perspective of people with illnesses or disabilities enjoying life in spite of their condition. It was normally identified through providing examples, rather than stating explicitly that enjoyment was a construct of quality of life.

You have people that have an enormously great quality of life who can't walk anywhere...because they have this great social structure and play cards all day. [Trialist, Australia]

On considering the item, informants were split between those who felt Enjoyment was important and relevant, and those who did not. For those who felt Enjoyment was not relevant, a motivating factor appeared to be that the item was too broad to be relevant.

What do you mean by enjoyment and pleasure?...I suppose not vague, but possibly ambiguous. [Health economist, UK]

EQ-5D-5L

Informants viewed the EQ-5D-5L measure as a simple and straight forward measure of health. The length and simplicity of the measure was viewed positively and a number of informants noted that the language used was appropriate.

I think the great beauty of this is that you can do this in two minutes flat. [Trialist, UK]

Informants viewed it as a measure of health state, which captured the determinants of health-related quality of life they had described previously. However informants noted that it did not capture the broad spectrum of quality of life they had described.

...that one is more broadly health. [Research doctor, Australia]

...it is not capturing how they feel about their life. It's, they are not saying "I have a good life" or not...This one is what you can do and what problems do you have. [Trialist, UK]

There was broad recognition of the usefulness of the measure in health research. This appeared to be motivated partly by awareness of a strong precedent of use of the EuroQol measure and its recognition by funding and rationing bodies.

It is hard to beat the EuroQol in terms of NICE guidance and everything that is out there already. [Health economist, Australia]

Informants who had previously used EQ-5D-3L thought that the increase in levels would improve the ability of the measure to record change in health state, reducing the "ceiling effect" which existed in the old measure and making it more attractive. Enthusiasm was shown for the new version.

I think I would prefer it to the EuroQol that we are using now. [Frontline researcher, Australia]

Mobility. In the prior discussion informants identified the ability to be mobile, as well as the ability to move upper and lower limbs as important determinants of quality of life. Mobility was not valued for itself, rather for allowing individuals the independence to access their normal everyday life.

...it is independence, it is to do with mobility, it's the getting to the shops, being able to do what you want to do, when you can do it... [Trialist, Australia]

Disagreement existed about whether the item fully assessed mobility. Informants noted that the item assessed a persons' ability to walk, not their ability to be independently mobile. This was considered to limit the scope of the item.

...that's just about walking, whereas people can be independently mobile in a wheel chair, they can actually have quite a high quality of life. [Trialist, Australia]

Self-care. Prior to viewing the measure, no informant directly identified self-care as an important determinant of quality of life.

Upon seeing the measure informants felt the Self-care item was narrow, and arbitrary. Many felt it should be considered as part of usual activities, rather than as a dimension in and of itself.

...as a sort of category of assessment...it is only one action, like making a cup of tea. It is a bit arbitrary really. [Trialist, UK]

Usual activities. Informants discussed the importance of people being able to complete normal activities, such as going to work and having social contact. In the broad conceptualisation of quality of life offered by informants in their prior discussions, a large number of the non-health determinants appeared to relate to usual activities.

...in terms of their participation, that they are not able to do things that they normally do...whether it is looking after the grandchildren or cooking meals or something like that. [Trialist, Australia]

Informants considered the Usual Activities item to be broad and noted the need for the clarifying statement. While there was hesitance in the language used referring to the breadth of the item, only a few informants directly stated that breadth was a problem.

Whereas usual activity, work, home work, leisure is massively broad. [Trialist, UK]

Pain/Discomfort. Pain was identified as a particularly important, almost pervasive, health-related influence on quality of life. Informants with clinical training discussed how pain could be managed to reduce its influence on quality of life.

Nothing worse for quality of life in many ways than chronic discomfort and pain. [Research doctor, Australia]

The Pain/Discomfort item was noted as being an important aspect to measure. Some concern existed about the phrasing of the question, with a small number of informants noting that the item assesses two distinct dimensions: pain or discomfort.

You got to wonder why you'd bother asking pain or discomfort. Wouldn't you ask one, because they are so different... [Trialist, Australia]

Anxiety/Depression. The effect of the psychological state upon quality of life was identified by a large number of informants and was thought to be influenced heavily by physical health. Worry, concern, fear, anxiety and depression were identified as important psychological determinants.

Depression is a frequent co-morbidity of severe physical illness. [Researcher doctor, Australia]

Informants thought the Anxiety/Depression item was very relevant. There was concern about the stigma attached to the word "depression" and how this might influence participants' answers. The use of the words "depression" and "anxiety" as a summary for psychological health was felt to lack scope.

...anxious and depressed...people don't like that word depression. You know, "don't tell me that"... [Frontline research, Australia]

Discussion

Informants considered the ICECAP-A to offer comprehensive measurement of the broad construct of quality of life they described, while lacking a direct assessment of health. The EQ-5D-5L was viewed as offering good coverage of health-related quality of life, while lacking assessment of the broader determinants of quality of life. This assessment is largely in line with the aims of each measure: the EQ-5D-5L is designed to measure the health determinants of quality of life, while the ICECAP-A capability measure's theoretical grounding focuses on wellbeing more broadly defined. These results therefore suggest that the ICECAP-A and EQ-5D-5L measures hold acceptable levels of content validity.

The item by item analysis showed that informants had concerns about some items in each measure. The content of the EQ-5D-5L Self-Care item was not viewed as relevant, while the restricted content of the Mobility item was questioned. In the ICECAP-A the content relevance of the Autonomy and Achievement items was thought to be age dependent. Other concerns pertained to the phrasing of some items (Achievement, Pain/Discomfort, Anxiety/Depression) or the potential for items to upset participants due to its subject matter (Attachment).

Both EQ-5D-5L and ICECAP-A were viewed as short, simple and easy to use, which is in line with findings from qualitative research with the general population for the EQ-5D-5L [7] and ICECAP-A [27]. The increase in levels of the EQ-5D-5L was expected to improve responsiveness and reduce "ceiling effects", this been shown in early quantitative assessments [11,12]. Concerns over interpretation of the capability wording in the ICECAP-A measure should be considered in light of research that found the general public were able to understand and answer the questions [27]. Use of the ICECAP-A in addition to the EQ-5D-5L rather than as a replacement was viewed as a positive step in assessing quality of life.

The rigorous qualitative methodology used is a notable strength of this study, which importantly adds an expert perspective to the validity portfolios of both measures. Although the number of informants interviewed was relatively small in absolute terms, importantly, it was sufficient to achieve saturation; there was however a slight oversampling of informants who work in cancer research. It was not possible to assess the effect that the informants familiarity with existing quality of life and health-related quality of life measures had on their opinions of these measures, this is particularly true for the findings of the EQ-5D-5L where a number of informants had used the original 3 level version. It was not possible to assess whether the order in which the measures were viewed influenced responses, however the random order of presentation should have controlled for this to some degree.

Further qualitative and quantitative research, providing assessments of the content validity, construct validity and responsiveness of both measures in different clinical areas will be important. In line with the objectives of this research, to assess the validity and acceptability of these measures in a health setting, the informants used were health research professionals. This may have led to an increased focus on the physical health determinants of quality of life. Both the EQ-5D-5L and ICECAP-A may be of use in social care, public health and mental health research. Further qualitative research examining the content validity of these measures with patients and researchers in these areas is highlighted as an important area of research.

In conclusion, although there are concerns about specific content of some individual items, this study offers evidence of the content validity and appropriateness of both measures in a trial context in the UK. Informants viewed the ICECAP-A as an assessment of the broader determinants of quality of life; while the EQ-5D-5L was viewed as offering good coverage of health determinants of quality of life. This is largely in line with the objectives of both measures. This research adds an expert perspective to the emerging validity portfolios of these measures and in doing so allows greater confidence in the validity of these measures.

Author Contributions

Conceived and designed the experiments: TK JC HAJ PL. Performed the experiments: TK. Analyzed the data: TK JC HAJ. Contributed reagents/materials/analysis tools: TK JC HAJ PL. Wrote the manuscript: TK JC HAJ PL.

References

1. Sen A (1980) Equality of what? In: A Sen. Choice, Welfare and Measurement. Oxford: Blackwell.
2. Sen A (1992) Inequality Reexamined. Oxford: Oxford University Press.
3. Sen A (1993) Capability and Well-Being. In: M NussbaumA Sen. The Quality of Life. Oxford: Oxford University Press. pp. 30-53.
4. Sen A (2002) Health: perception versus observation. BMJ 324: 860-861. doi:10.1136/bmj.324.7342.860. PubMed: 11950717.
5. Coast J, Smith RD, Lorgelly P (2008) Welfarism, extra-welfarism and capability: the spread of ideas in health economics. Soc Sci Med 67: 1190-1198. doi:10.1016/j.socscimed.2008.06.027. PubMed: 18657346.
6. Al-Janabi H, Flynn TN, Coast J (2012) Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. Qual Life Res 21: 167-176. doi:10.1007/s11136-011-9927-2. PubMed: 21598064.
7. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P et al. (2011) Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research 20: 1727-1736. doi: 10.1007/s11136-011-9903-x. PubMed: 21479777.
8. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ (2008) Comparing the standard EQ-5D three-level system with a five-level version. Value Health 11: 275-284. doi:10.1111/j.1524-4733.2007.00230.x. PubMed: 18380640.
9. Janssen MF, Birnie E, Bonsel GJ (2008) Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. Qual Life Res 17: 463-473. doi: 10.1007/s11136-008-9318-5. PubMed: 18320352.
10. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S (2007) Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care 45: 259-263. doi:10.1097/01.mlr.0000254515.63841.81. PubMed: 17304084.
11. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M 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: 1717–27. doi:10.1007/s11136-012-0322-4. PubMed: 23184421.
12. Kim SH, Jo MW (2011) Psychometric Comparision of the Eq-5D-3l to the Eq-5D-5l in Cancer Patients in Korea. Value in Health 14: A171. doi:10.1016/j.jval.2011.02.946.
13. EuroQol (2012). Available: http://www.euroqol.org/eq-5d-products/eq-5d-5l.html.
14. Magasi S, Ryan G, Revicki D, Lenderking W, Hays RD et al. (2012) Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res 21: 739-746. doi:10.1007/s11136-011-9990-8. PubMed: 21866374.
15. Cronbach LJ, Meehl PE (1955) Construct Validity in Psychological Tests. Psychol Bull 52: 281-302. doi:10.1037/h0040957. PubMed: 13245896.
16. Carmines EG, Zeller RA (1979) Reliability and Validity Assessment. London: Sage Publications.
17. Brod M, Tesler LE, Christensen TL (2009) Qualitative research and content validity: developing best practices based on science and experience. Qual Life Res 18: 1263-1278. doi:10.1007/s11136-009-9540-9. PubMed: 19784865.
18. Sireci SG (1998) The construct of content validity. Social Indicators Research 45: 83-117. doi:10.1023/A:1006985528729.
19. Lasch KE, Marquis P, Vigneux M, Abetz L, Arnould B et al. (2010) PRO development: rigorous qualitative research as the crucial foundation.

Qual Life Res 19: 1087-1096. doi:10.1007/s11136-010-9677-6. PubMed: 20512662.
20. Kim TH, Jo MW, Lee SI, Kim SH, Chung SM (2013) Psychometric properties of the EQ-5D-5L in the general population of South Korea. Qual Life Res, 22: 2245–53. doi:10.1007/s11136-012-0331-3. PubMed: 23224560.
21. Scalone L, Ciampichini R, Fagiuoli S, Gardini I, Fusco F et al. (2013) Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. Qual Life Res, 22: 1707–16. doi:10.1007/s11136-012-0318-0. PubMed: 23192232.
22. Al-Janabi H, Peters TJ, Brazier J, Bryan S, Flynn TN et al. (2013) An investigation of the construct validity of the ICECAP-A capability measure. Qual Life Res, 22: 1831–40. doi:10.1007/s11136-012-0293-5. PubMed: 23086535.
23. Ritchie J, Lewis J, Elam G (2009) Designing ans Selecting Samples. In: J RitchieJ Lewis. Qualitative Research Practice. London: SAGE Publications. pp. 77-108.
24. Legard R, Keegan J, Ward K (2009) In-depth Interviews. In: J RitchieJ Lewis. Qualitative Research Practice. London: SAGE Publishing. pp. 138-169.
25. Spencer L, Ritchie J, O'Connor W (2009) Analysis: Practices, Principles and Processes. In: J RitchieJ Lewis. Qualitative Research Practice. London: SAGE Publications. pp. 199-218.
26. Ritchie J, Spencer L, O'Connor W (2009) Carrying out Qualitative Analysis. In: J RitchieJ Lewis. Qualitative Research Practice. London: Sage. pp. 219-262.
27. Al-Janabi H, Keeley T, Mitchell P, Coast J (2013) Can capabilities be self-reported? A think aloud study. Social Science and Medicine 87: 116-122.

Appendix 11

Trial recruitment letter

Dear ,

My name is Tom Keeley, I am a PhD student in the University's MRC Trials Methodology Hub. <Name> has suggested that I contact you to discuss the possibility of incorporating a new, short, outcome measure (the ICECAP capabilities index) for use in cost-effectiveness analysis into a trial within the <name of trials unit>

The ICECAP measure may allow for a broader assessment of improvements in quality of life following health interventions than is available from using the current EQ-5D. It has the same sort of properties as the QALY approach in that it is possible to utilise the measure across interventions that impact on both quality and quantity of life. The measure has five dimensions: Attachment (love and friendship), Security (thinking about the future without concern), Enjoyment (enjoyment and pleasure), role (doing things that make you feel valued) and control (independence). It has already been shown to have validity in cross-sectional studies and sensitivity to change in the context of total joint replacement. As a practical measure it is comparable in length and simplicity to the EQ-5D – like the EQ-5D, it only takes one page of A4 and has only five questions. There is more information about the measure and its development on its website: http://www.icecap.bham.ac.uk/ICECAP-A/index.shtml.

My PhD project is part of a continuing research stream within the Health Economics Unit into the use of the ICECAP measure, and is particularly concerned with looking at the implications for cost-effectiveness of using this different approach to economic evaluation. My aim is to compare assessments of cost-effectiveness using the ICECAP measure with those obtained using other measures such as QALYs obtained from EQ-5D or SF-6D, so it would be especially helpful to be able to put the ICECAP measure alongside EQ-5D or SF-36/SF-12 in studies where you were already planning to include one or other of these measures.

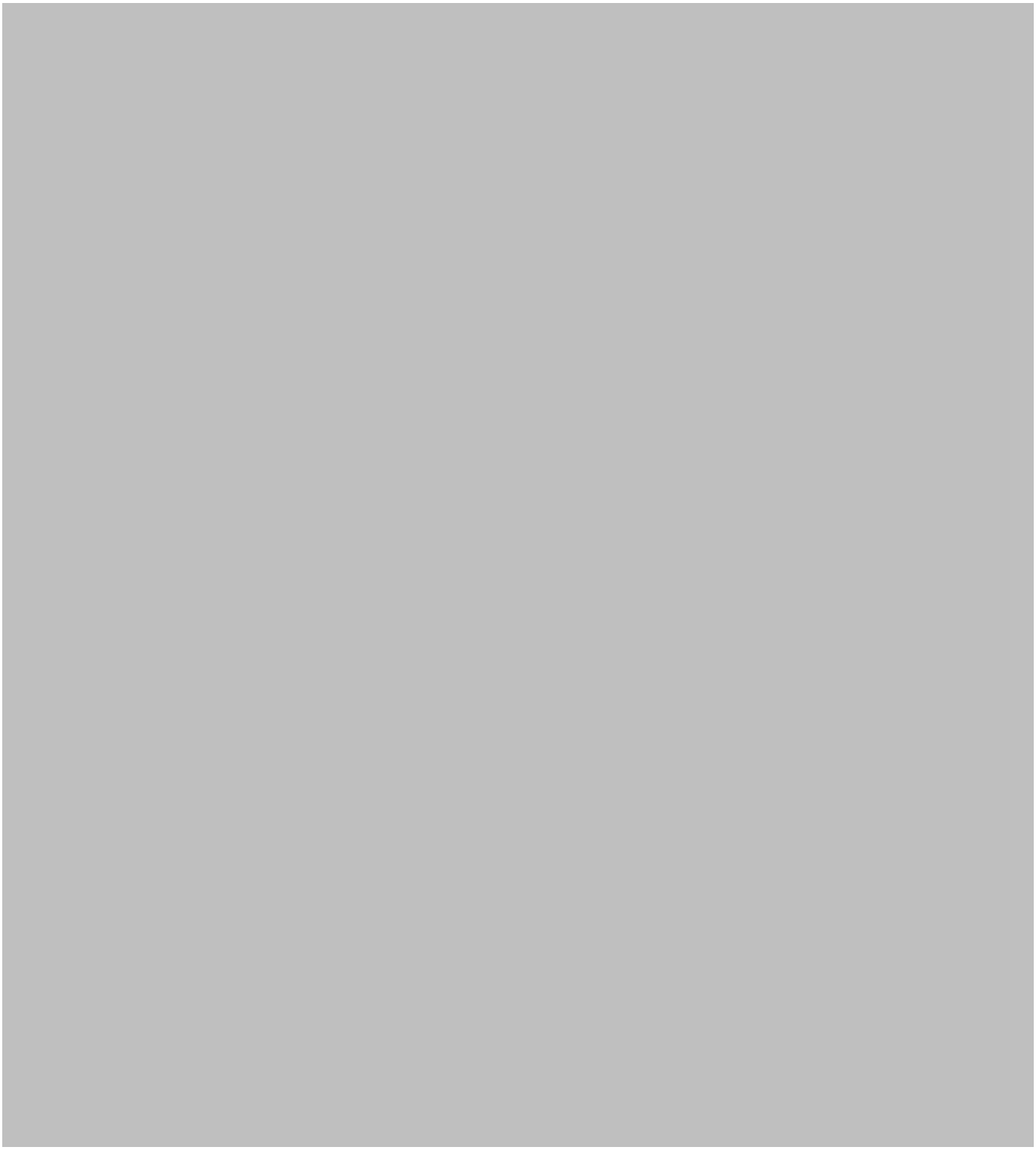
I would really appreciate it if my supervisors (Jo Coast, Hareth Al-Janabi) and I could meet with you in the near future to talk about the possibility of including this measure within a trial in your unit, and hope that you won't mind if I ring in a week or so to try to arrange this. Please feel free to get back to me with any questions about the measure or the work that I am doing.

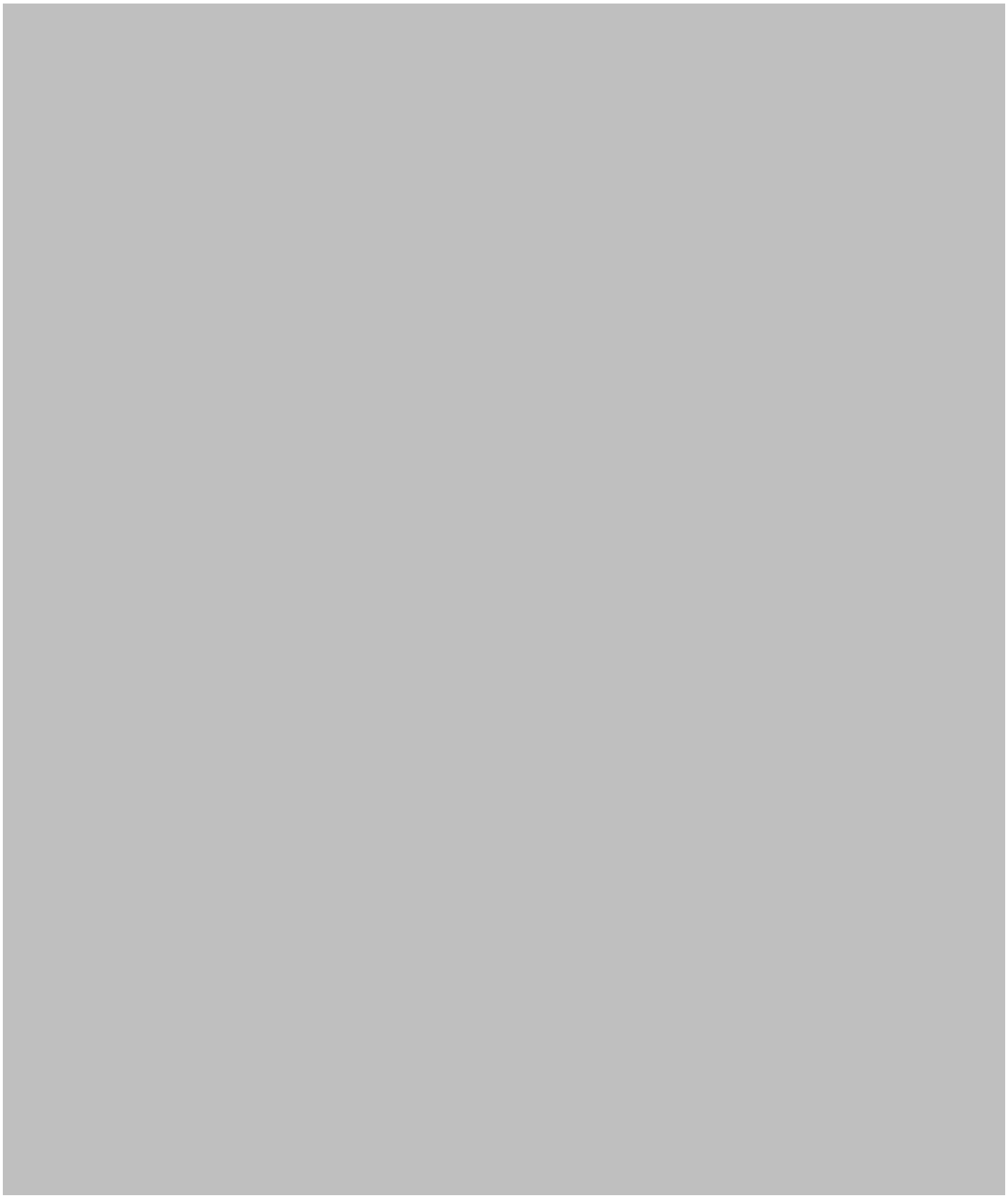
With best wishes,

Tom

Appendix 12

EQ-5D-3L



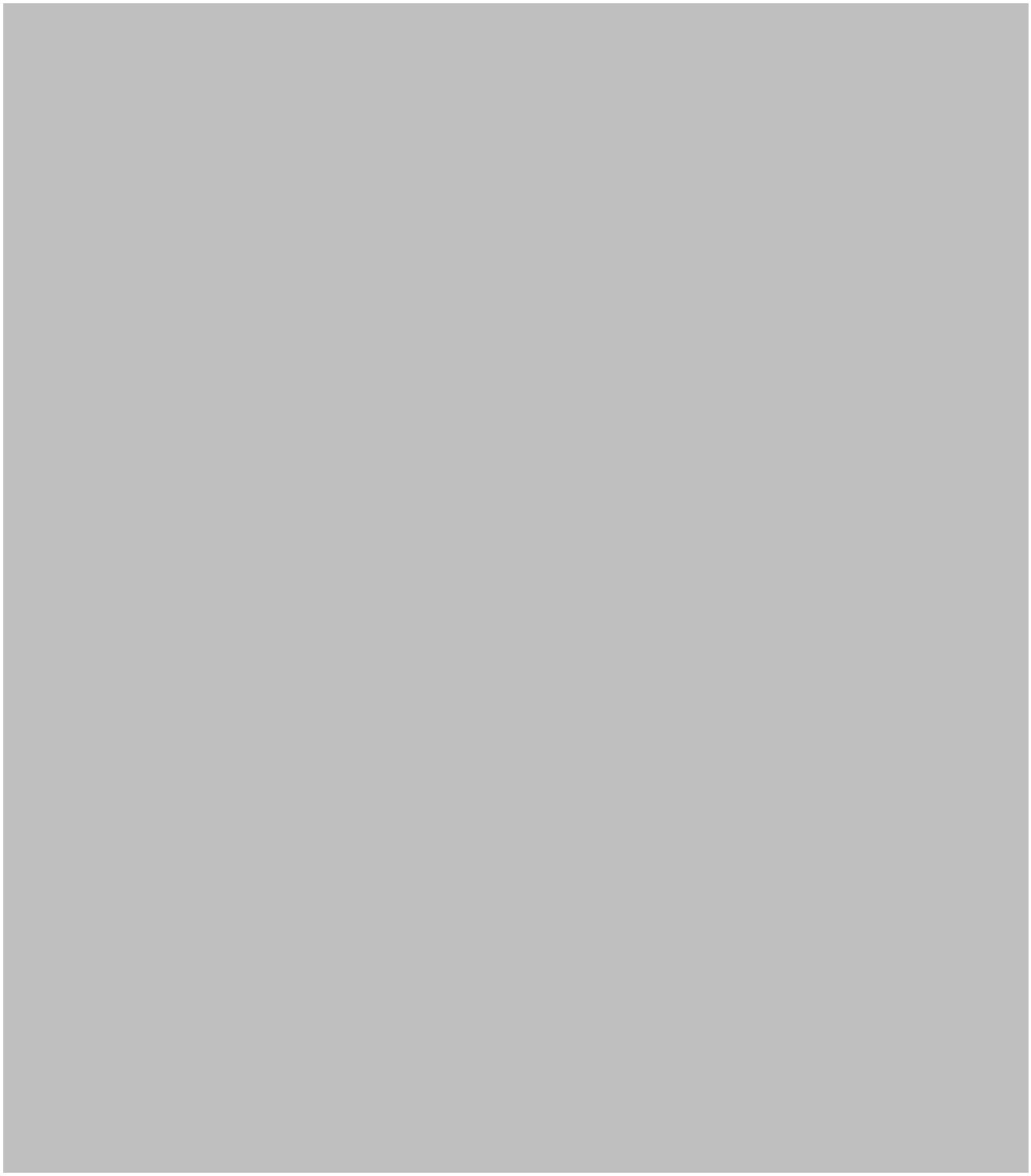


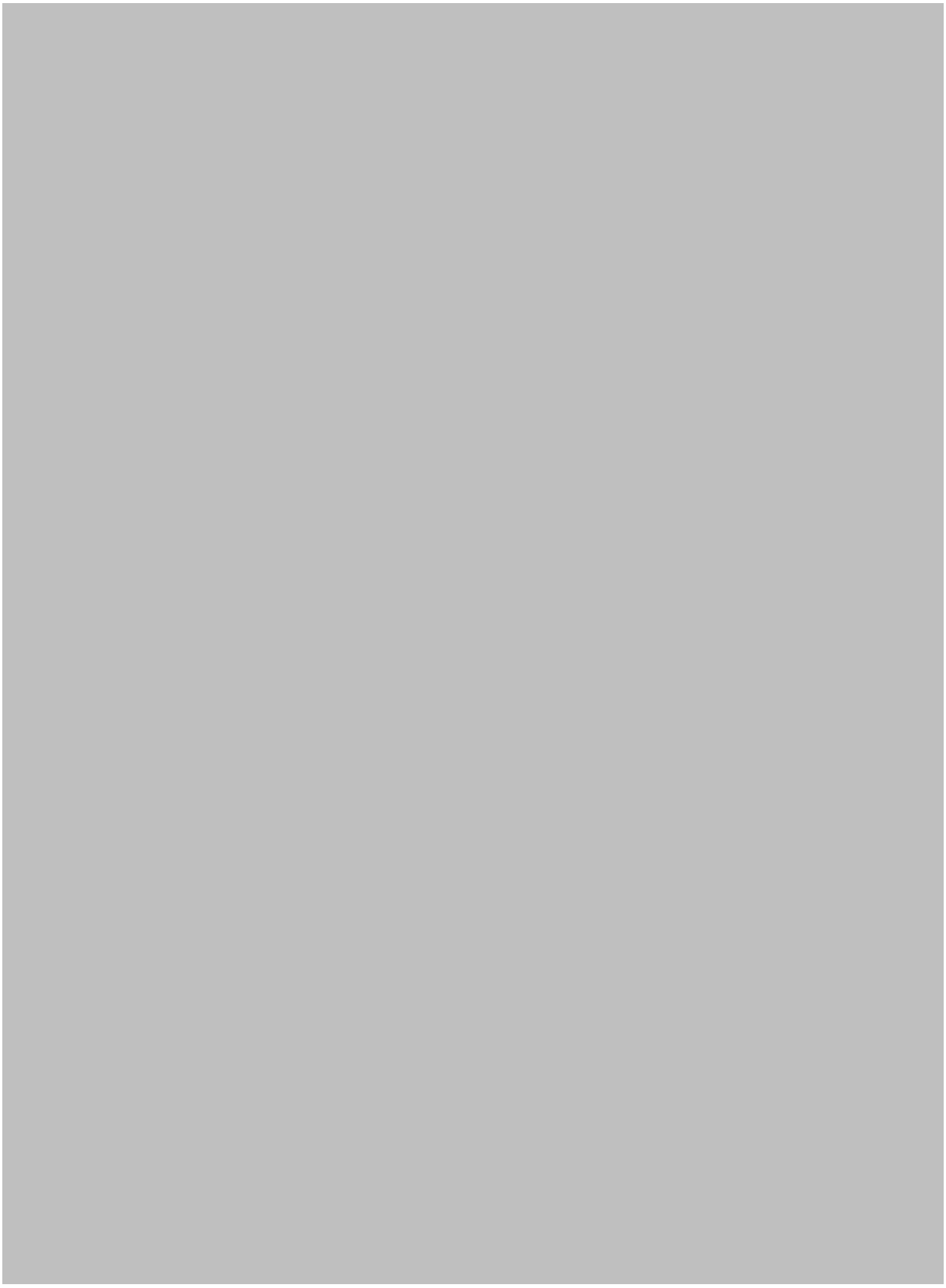
Appendix 13

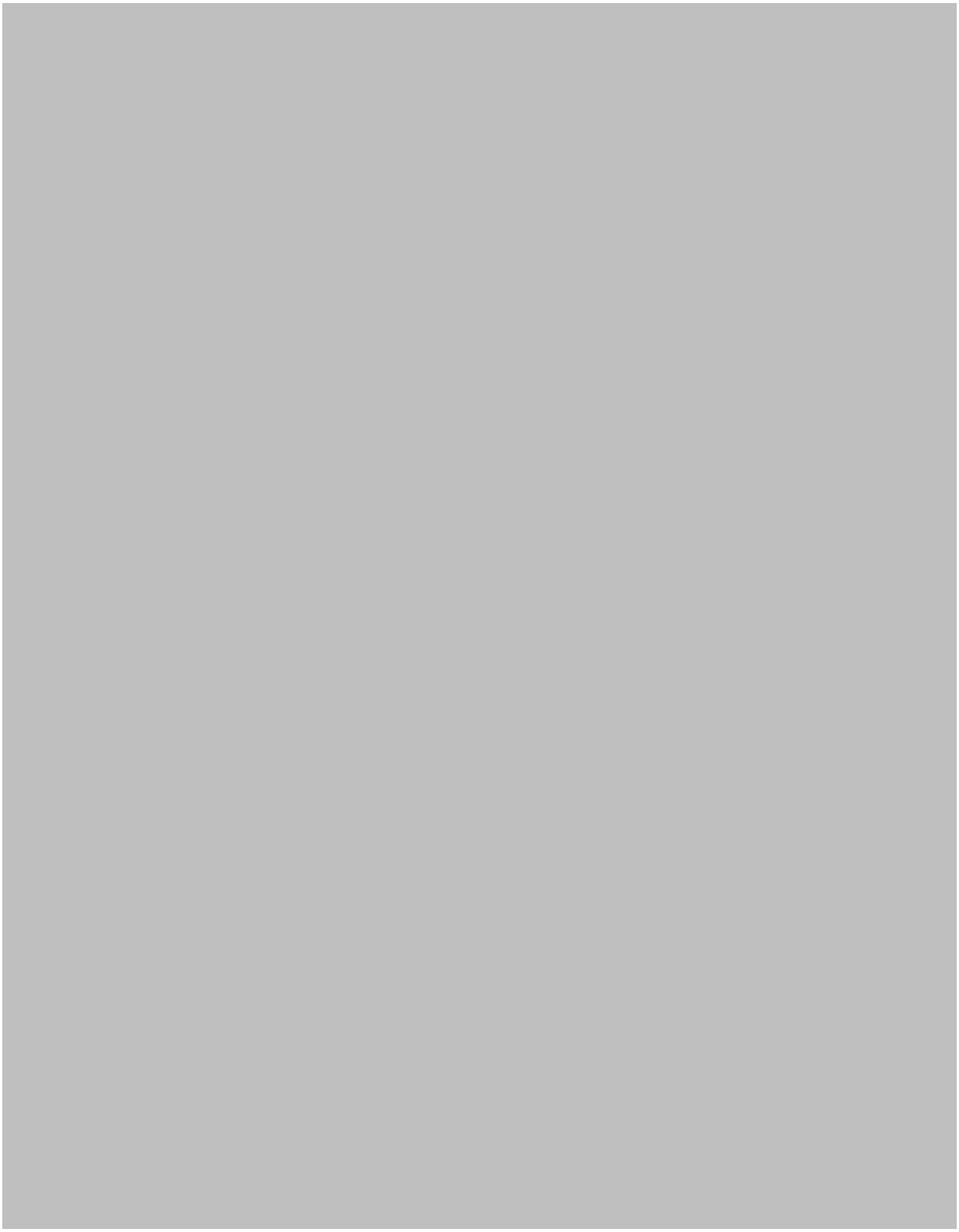
SF-36

SF-36 QUESTIONNAIRE

(1992 -- Medical Outcomes Trust)







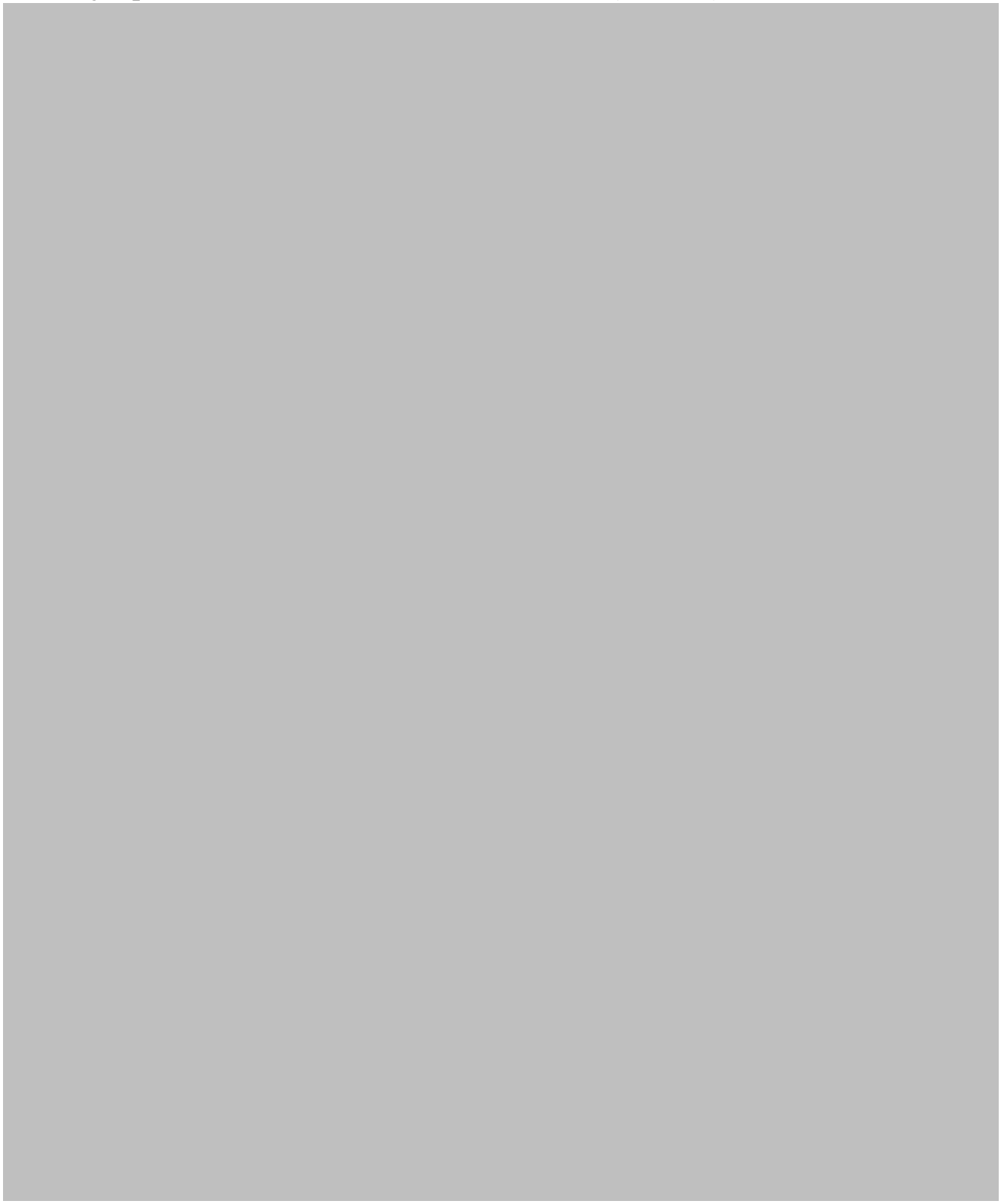
Appendix 14

Co-morbidities measurement (BEEP)

t 4

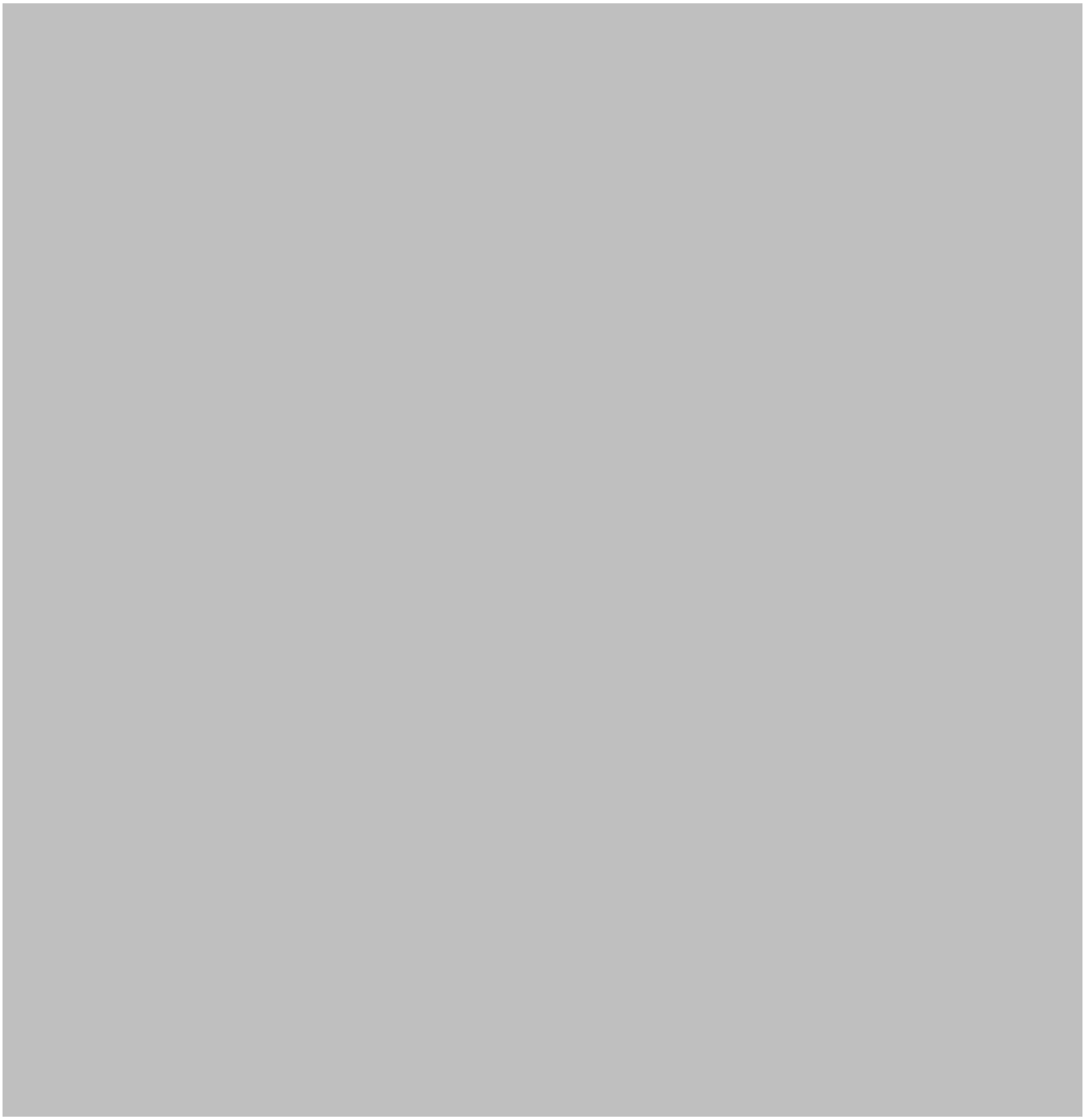
Appendix 15

Symptoms and side effects measurement (PastBP)



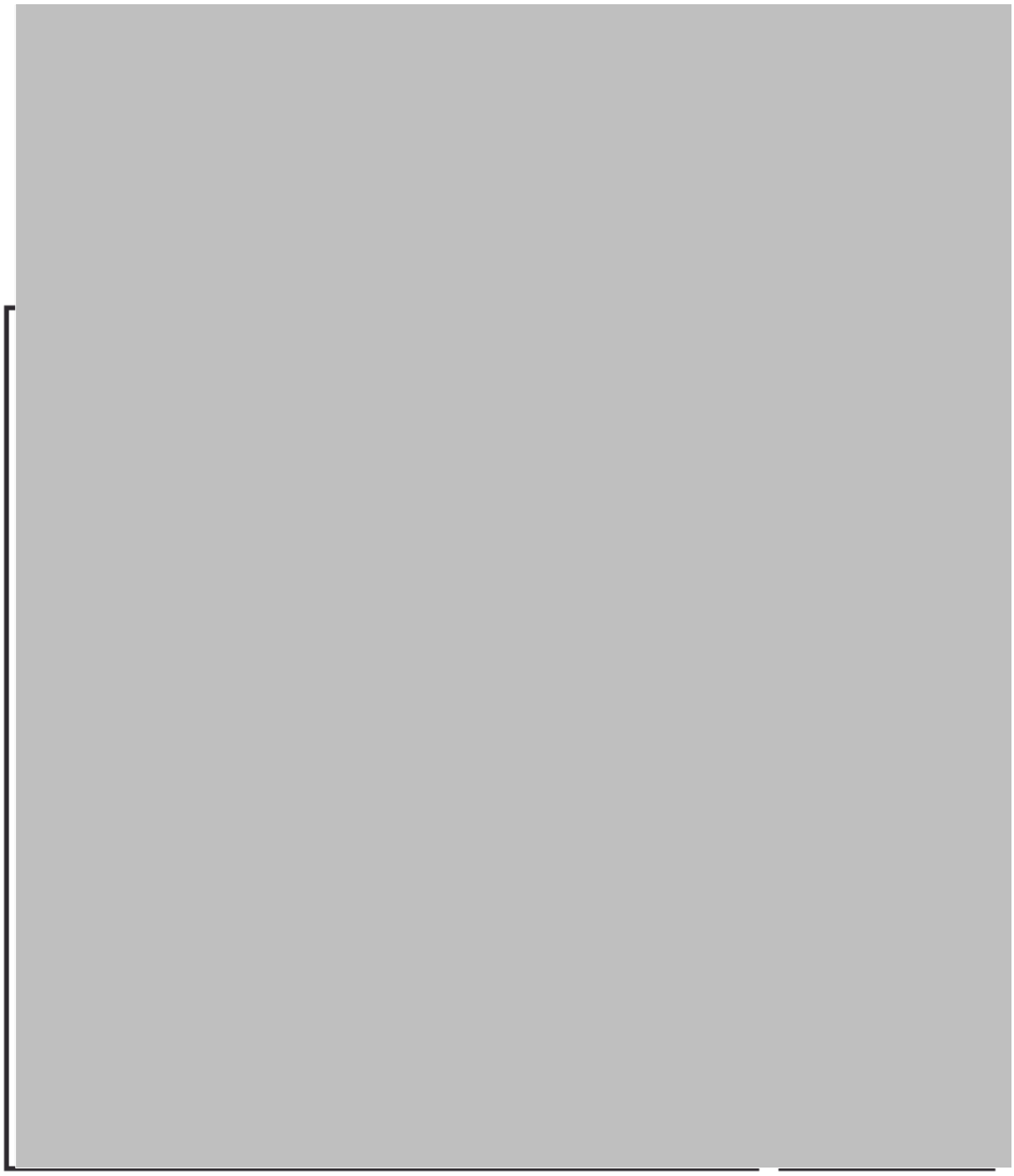
Appendix 16
WOMAC

WOMAC Osteoarthritis Index LK3.1 (IK)

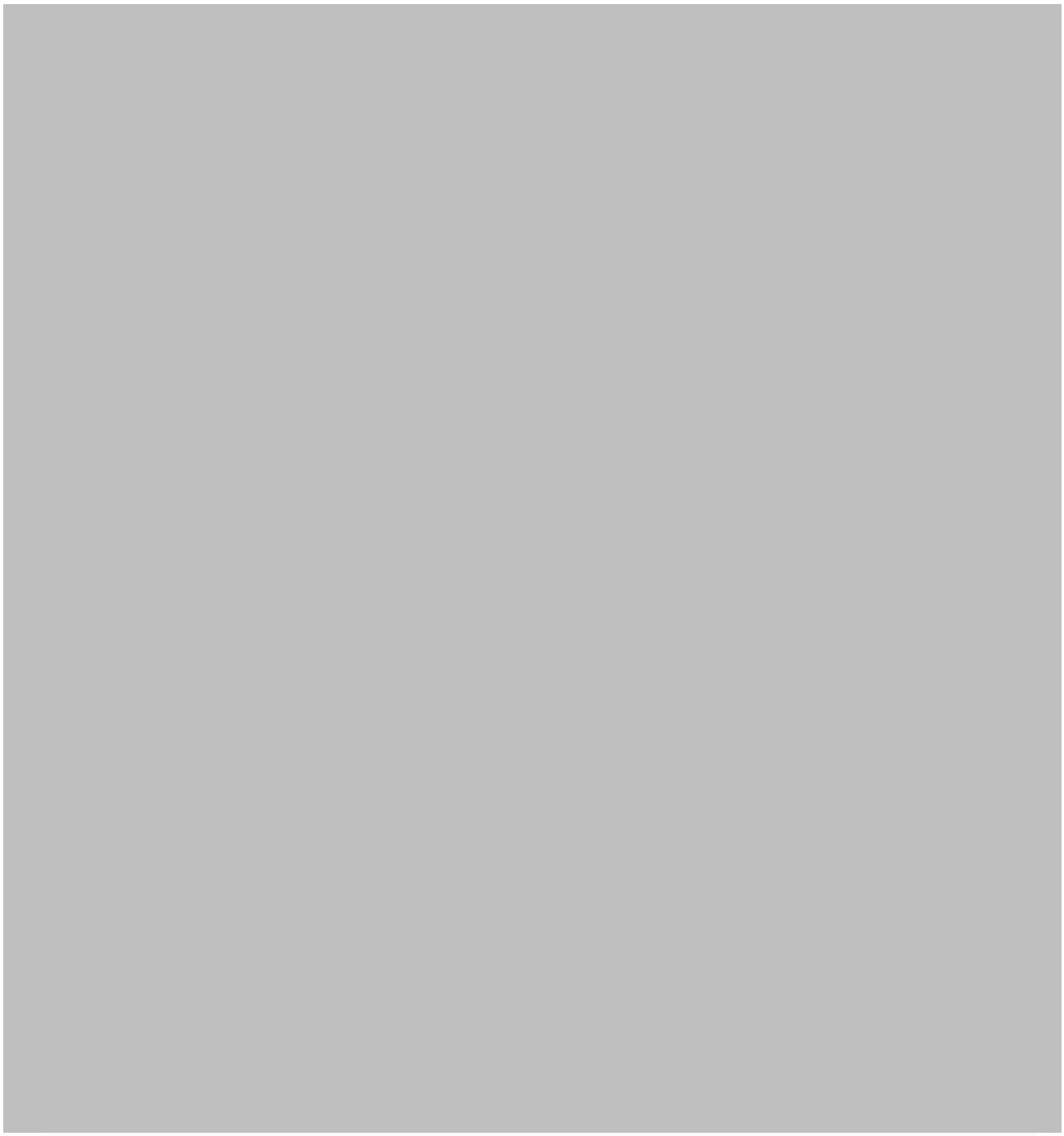


WOMAC Osteoarthritis Index LK3.1 (IK)

WOMAC Osteoarthritis Index LK3.1 (IK)



WOMAC Osteoarthritis Index LK3.1 (IK)

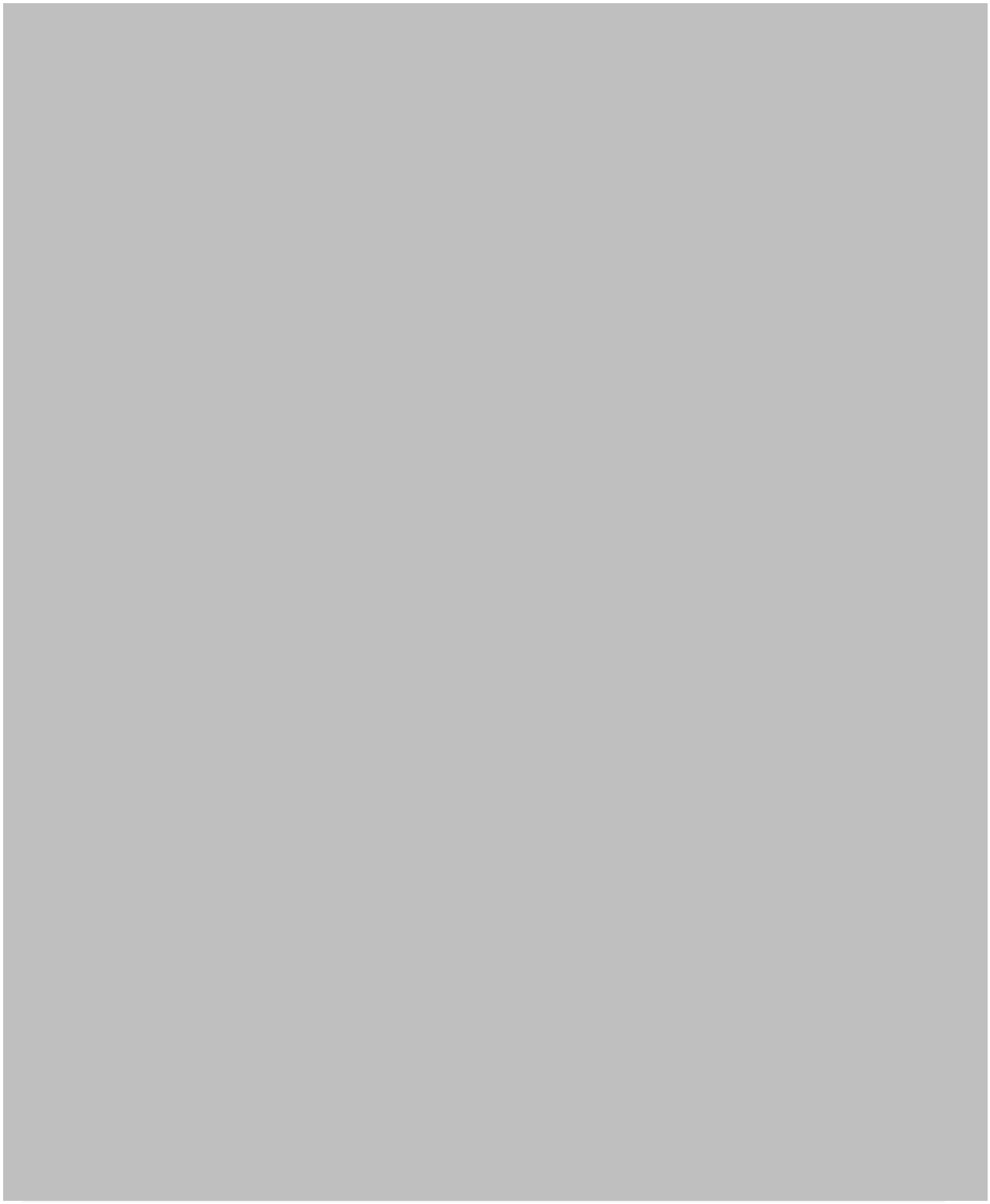


WOMAC Osteoarthritis Index LK3.1 (IK)

Appendix 17

Brief Illness Perception Questionnaire

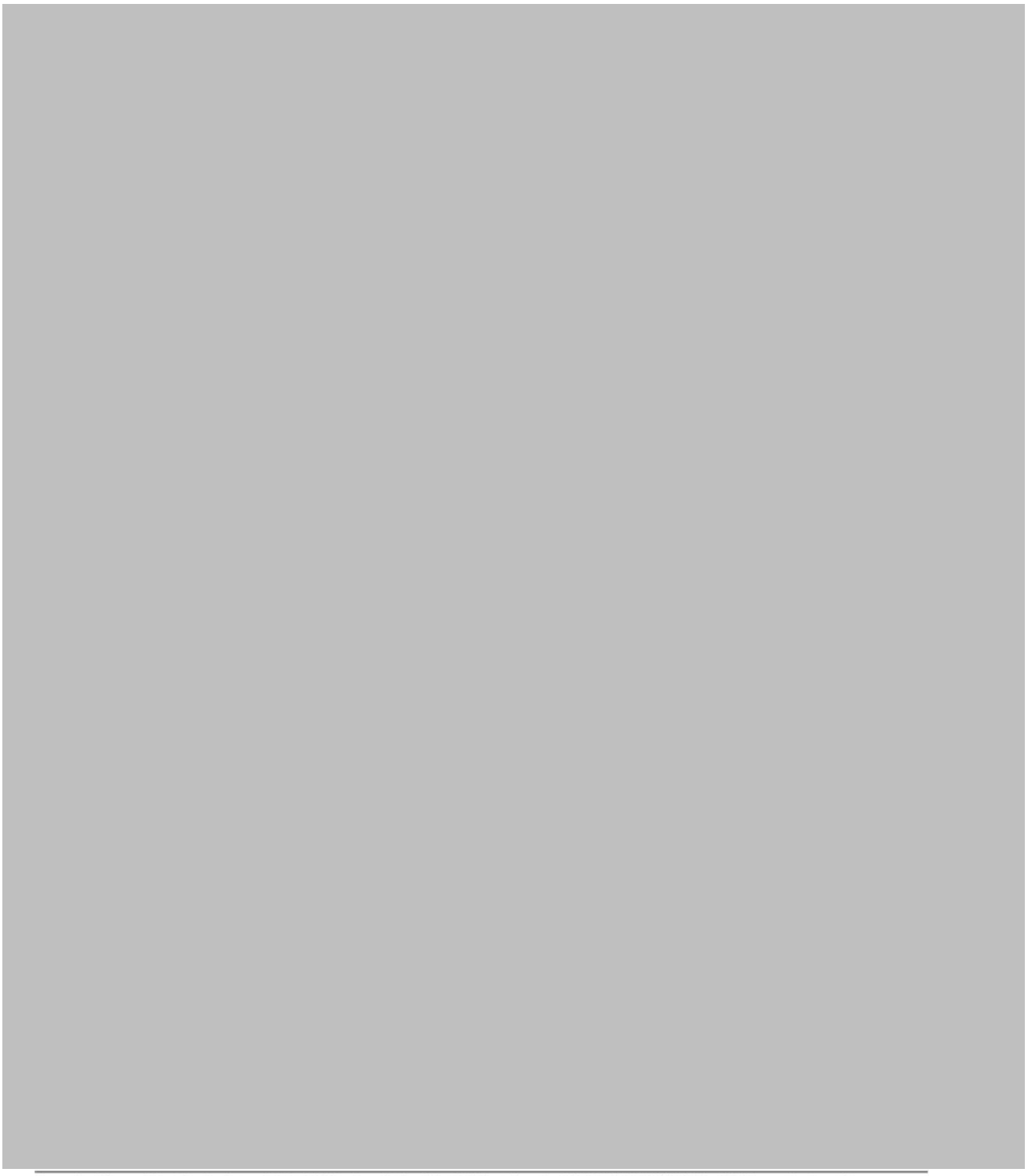
The Brief Illness Perception Questionnaire



Appendix 18

GAD-7 and PHQ-8

Generalized Anxiety Disorder 7-item (GAD-7) scale



From the Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MDPHQ). The PHQ was developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues. For research in formation, contact Dr. Spitzera trls8@columbia.edu. PRIME-MD® is a trademark of Pfizer Inc. Copyright© 1999 Pfizer Inc. All rights reserved. Reproduced with permission

Appendix 19

Hypotheses for BEEP trial

Expected association between the ICECAP-A tariff score and the socio-demographic, physical health and psychological health variables

	Hypotheses	
	Association	Direction
Socio-demographic		
Age	Yes	Negative
Gender	No	
Physical health		
EQ-5D-3L	Yes	Positive
WOMAC Pain	Yes	Negative
WOMAC stiffness	Yes	Negative
WOMAC functioning	Yes	Negative
IPQ		*Exploratory analysis*
Co-morbidities		*Exploratory analysis*
Psychological health		
GAD-7	Yes	Negative
PHQ-8	Yes	Negative

* Strength refers to the strength of a correlations between the variable and the ICECAP- tariff score.

Expected associations between gender and the ICECAP-A items

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
Gender	No	No	No	No	No

Expected associations between age and ICECAP-A items

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
Age	No	No	Yes	No	No

Expected associations between ICECAP-A and EQ-5D-3L items

Comparator	Mobility	Self-care	Usual Activities	Pain/ Discomfort	Anxiety/ depression
ICECAP-A	Yes	Yes	Yes	Yes	Yes

Expected associations between EQ-5D-3L tariff score and ICECAP-A items

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
EQ-5D-3L	Yes	No	Yes	Yes	Yes

Expected direction and strength of associations between EQ-5D-3L items and the ICECAP-A items

Comparator variables		Stability	Attachment	Autonomy	Achievement	Enjoyment
Mobility	Association	Yes		Yes	Yes	Yes
	Strength			Strong*		
Self-care	Association	Yes		Yes	Yes	Yes
	Strength			Strong*		
Usual activities	Association			Yes	Yes	Yes
	Strength			Strong*		
Pain/ discomfort	Association	Yes		Yes	Yes	Yes
	Strength					Strong*
Anxiety/ depression	Association	Yes		Yes	Yes	Yes
	Strength					Strong*

* Researchers were asked to predict the ICECAP-A item that would show the strongest association with each EQ-5D-3L item. Therefore "strong" is not a numerical prediction of a correlation rather prediction of the strength of the correlation in relation to other ICECAP-A items for that EQ-5D-3L item.

Expected associations between WOMAC sub-scales and ICECAP-A items

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
WOMAC Pain	Yes		Yes	Yes	Yes
WOMAC stiffness	Yes		Yes	Yes	Yes
WOMAC functioning	Yes		Yes	Yes	Yes

Expected associations between measures of anxiety and depression and ICECAP-A items

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
GAD-7	Yes		Yes		Yes
PHQ-8	Yes		Yes		Yes

Appendix 20

Hypotheses for PastBP trial

Expected associations between the ICECAP-O tariff score and socio-demographic, physical and psychological variables

	Hypotheses	
	Association	Direction
Socio-demographic		
Age	Yes	Negative
Gender	No	
Physical health		
EQ-5D-3L	Yes	Positive
EQ-5D-3L VAS	Yes	Positive
SSE	*Exploratory analysis*	
Physical functioning	Yes	Positive
Physical role	Yes	Positive
Bodily pain	Yes	Negative
General heatlh	Yes	Positive
Psychological health		
Vitality	Yes	Positive
Social functioning	Yes	Positive
Emotional role	Yes	Positive
Mental health	Yes	Positive

Expected associations between gender and the ICECAP-0 items

Comparator	Attachment	Security	Role	Enjoyment	Control
Gender	No	No	No	No	No

Expected associations between age and ICECAP-O items

Comparator	Attachment	Security	Role	Enjoyment	Control
Age	No	No	No	No	Yes

Expected associations between ICECAP-O and EQ-5D-3L items

Comparator	Mobility	Self-care	Usual Activities	Pain/ Discomfort	Anxiety/ depression
ICECAP-O	Yes	Yes	Yes	Yes	Yes

Expected associations between EQ-5D-3L tariff score and ICECAP-O items

Comparator	Attachment	Security	Role	Enjoyment	Control
EQ-5D-3L	No	Yes	Yes	Yes	Yes

Expected strength of associations between EQ-5D-3L items and the ICECAP-O items

Comparator variables		Attachment	Security	Role	Enjoyment	Control
Mobility	Association			Yes	Yes	Yes
	Strength					Strong*
Self-care	Association		Yes	Yes	Yes	Yes
	Strength					Strong*
Usual activities	Association			Yes	Yes	Yes
	Strength			Strong*		Strong*
Pain/ discomfort	Association		Yes	Yes	Yes	Yes
	Strength				Strong*	
Anxiety/ depression	Association		Yes	Yes	Yes	Yes
	Strength				Strong*	

* Researchers were asked to predicted the ICECAP-O item that would show the strongest association with each EQ-5D-3L item. Therefore "strong" is not a numerical prediction of a correlation rather prediction of the strength of the correlation in relation to other ICECAP-O items for that EQ-5D-3L item. In the case of usual activities no agreement was reached on which ICECAP-O item would associated strongest, therefore both Role and Control were marked as expected a strong association.

Expected strength of associations between SF-36 subscales and ICECAP-O tariff scores

SF-36		Correlation with ICECAP-O tariff
Physical		
Physical functioning	Association	Yes
	Strength	
Physical role	Association	Yes
	Strength	
Bodily Pain	Association	Yes
	Strength	
General Health	Association	Yes
	Strength	Strong
Psychological		
Vitality	Association	Yes
	Strength	Strong
Social functioning	Association	Yes
	Strength	Strong
Emotional role	Association	Yes
	Strength	
Mental health	Association	Yes
	Strength	

* Researchers were asked to predict which SF-36 sub-scales would have the strongest association with the ICECAP-O tariff. Therefore "strong" is not a numerical prediction of a correlation rather prediction of the strength of the correlation in relation to other ICECAP-O /SF-36 sub-scale correlations.

Expected associations between SF-36 sub-scales and ICECAP-O items

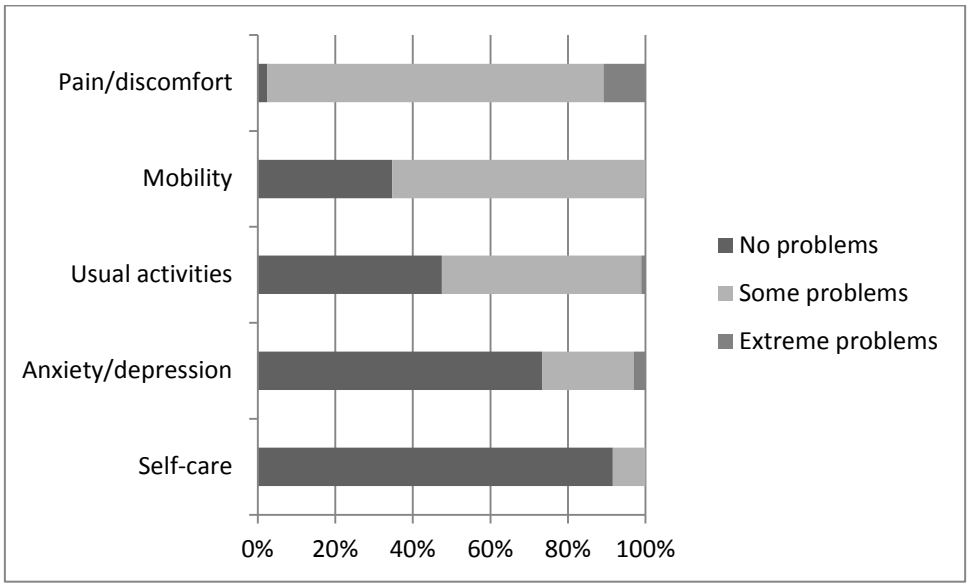
SF-36	Attachment	Security	Role	Enjoyment	Control
Physical					
Phys func		Yes	Yes	Yes	Yes
Phys role		Yes	Yes	Yes	Yes
Pain		Yes	Yes	Yes	Yes
Gen health		Yes	Yes	Yes	Yes
Psychological					
Vitality			Yes	Yes	Yes
Social func	Yes		Yes	Yes	Yes
Emo role	Yes		Yes	Yes	Yes
Mental health		Yes	Yes	Yes	Yes

Expected associations between Modified Rankin Scale and ICECAP-O items

Comparator	Attachment	Security	Role	Enjoyment	Control
MRS	No	Yes	Yes	No	Yes

Appendix 21

Detailed breakdown of baseline EQ-5D-3L scores (BEEP)



Appendix 22

Brief Illness Perception Questionnaire Analysis

The analysis of the Brief IPQ was exploratory. Little information on the development or validity of the measure could be found on which to base expectations of the association with the ICECAP-A. Furthermore, no previous studies had assessed the association of the ICECAP-A with this measure, or any measure like it. Therefore, no *a priori* hypotheses were formed.

As can be seen from Table A the overall Brief IPQ score showed a weak, but statistically significant correlation with the ICECAP-A tariff score. The Brief IPQ is made up of 8 questions. Each of the 8 questions has 11 response options (a 0-10 scale). Therefore each of the variables is an 11 level categorical variable. To avoid low numbers each variable was recoded to form a 3 level categorical variable of low (0-3) medium (4-6) and high (7-10). Four Variables showed a significant association with the ICECAP tariff score at the 5% significance level: Affect, an assessment of the affect that knee pain has on life; symptoms, an assessment of how many symptoms are experienced; concerned, an assessment of the level of concern over knee pain; and emotion, an assessment of the emotional effect of knee pain.

Table A: Correlationss between ICECAP-A tariff score and IPQ questions

	Affect	Continue	Control	Treatment	Symptoms	Concerned	Understand	Emotion
ICECAP A tariff	<0.001 **	0.233	0.77	0.4	0.023*	0.016*	0.902	<0.001 **

Table B shows the associations between each item of the Brief IPQ and the ICECAP-A items. There are a number of significant associations at the 5% and 1% significance level. Affect,

Symptoms and Emotion were most likely to associate with the Brief IPQ items. Attachment showed association with Emotion at the 5% level of significance. It can therefore be inferred from these results that a person's perception of the affect knee pain has upon their life, the frequency of symptoms experienced and the emotional affect of knee pain is reflected in the scores of the ICECAP-A measure.

Table B: Associations between ICECAP-A items and IPQ questions

Comparator	Stability	Attachment	Autonomy	Achievement	Enjoyment
Affect	0.029**	0.222	<0.001**	<0.001**	<0.001**
Continue	0.798	0.960	0.284	0.248	0.223
Control	0.985	0.658	0.952	0.634	0.377
Treatment	0.990	0.834	0.197	0.171	0.587
Symptoms	0.002**	0.103	0.02*	<0.001**	0.017*
Concerned	0.040*	0.442	0.336	0.145*	0.027**
Understand	0.549	0.910	0.711	0.368	0.688
Emotionally	<0.001**	0.023*	<0.001**	<0.001**	<0.001**

Appendix 23

ICECAP-A response profile for worsened EQ-5D-3L index scores

	Baseline profile					Follow-up profile					Change between baseline and follow-up				
	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy
Level 1	0	0	0	0	0	3	0	0	0	0	+3	0	0	0	0
Level 2	10	5	3	13	8	18	13	5	18	18	+8	+8	+2	+5	+10
Level 3	58	29	37	63	45	58	29	55	63	55	0	0	+18	0	+10
Level 4	32	66	60	24	47	21	58	39	18	26	-11	-8	-21	-6	-21

Appendix 24

ICECAP-A response profiles for improved EQ-5D-3L index scores

	Baseline profile					12 month follow-up profile					Change between baseline and follow-up				
	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy
Level 1	4	0	0	0	0	1	1	1	1	1	+1	+1	+1	+1	+1
Level 2	15	8	3	13	10	9	5	4	10	9	-6	-3	+1	+3	-1
Level 3	55	33	43	62	52	56	31	28	54	50	+1	-2	-15	-8	-2
Level 4	26	59	54	25	38	34	63	67	35	39	+8	+4	+13	+10	+1

Appendix 25

ICECAP-A response profile for worsened GAD-7 health status

	Baseline profile					12 month follow-up profile					Change between baseline and follow-up				
	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy
Level 1	5	0	0	0	0	7	0	2	5	5	+2	0	+2	+5	+5
Level 2	11	9	7	12	12	30	21	7	16	16	+19	+12	0	+4	+4
Level 3	58	30	33	65	53	49	33	30	65	58	-9	+3	-3	0	+5
Level 4	26	60	60	23	35	14	46	60	14	21	-12	-14	0	-9	-14

Appendix 26

ICECAP-A response profile for improved GAD-7 health status

	Baseline profile					12 month follow-up profile					Change between baseline and follow-up				
	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy
Level 1	4	0	0	0	0	0	1	0	0	1	-4	+1	0	0	+1
Level 2	19	11	5	17	16	18	8	2	12	12	-1	-3	-3	-5	-4
Level 3	59	32	38	65	60	55	32	36	69	57	-4	0	-2	+4	-3
Level 4	18	57	57	18	24	27	58	61	19	30	+9	+1	+4	+1	+6

Appendix 27

PHQ-8 Anchor analysis

Anchor groups from the PHQ-8 score were not formed using minimally important changes. No values could be found in existing literature. The interquartile range of change was used and groups were formed based on a change of 2. No assumption can be made as to whether a change of 2 is important to participants. The mean change in PHQ-8 scores in these groups was significantly larger than 2 and these values are presented in Table A

Table A: Group numbers and mean PHQ-8 change scores in the PHQ-8 anchor groups (n=331)

Anchor group	Number in group	Mean PHQ-8 change in group (95% CI)	PHQ-8 change as a % of possible change
Improved	92	-4.473 (-5.139, -3.807)	18.6%
No change	185	-0.08 (-0.176, 0.016)	0.3%
Worsened	54	4.254 (3.455, 5.053)	17.7%

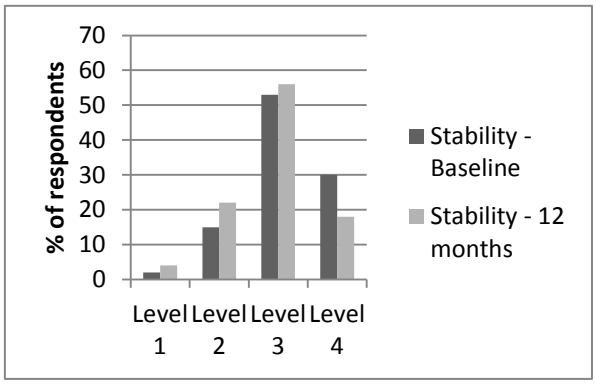
Item by Item analysis

Figure A show the response profiles at baseline and follow-up the group of respondents reporting a worsening of their PHQ-8 psychological health status (increase in PHQ-8 scores). These results are presented in numerical form in Appendix 28. The response profile of these participants changed between baseline and follow-up. A reduction of 23 points in the percentage of respondents answering the top level (level 4) of Enjoyment was found.

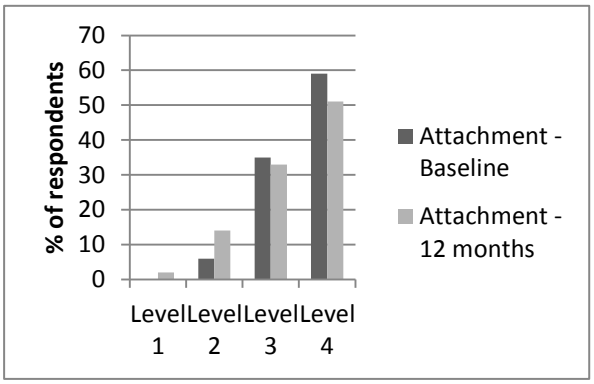
Smaller reductions of 8 points for Attachment and Achievement and 12 points for Stability were seen in the percentage of people answering level 4. Little change in the response profile of Autonomy was found.

Figure A: ICECAP-A response profile at baseline and follow-up for participants reporting a worsening of their PHQ-8 health status

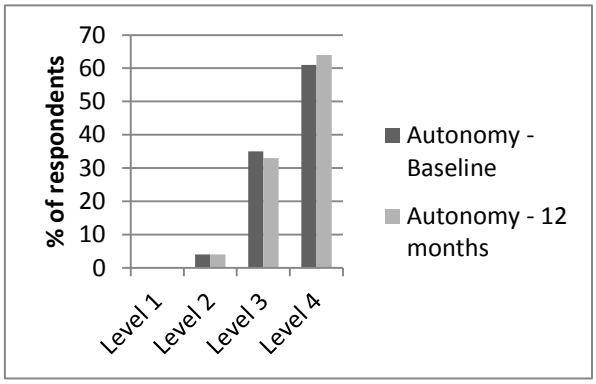
Stability Item



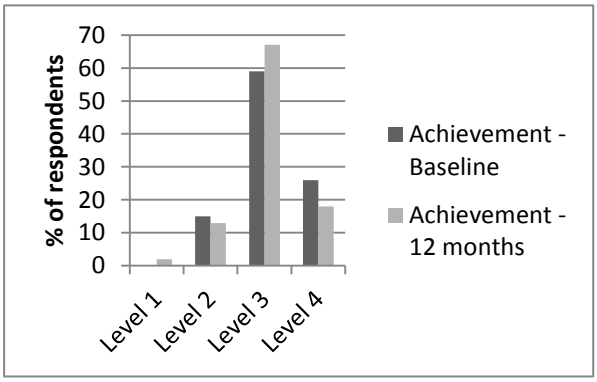
Attachment Item



Autonomy Item



Achievement Item



Enjoyment Item

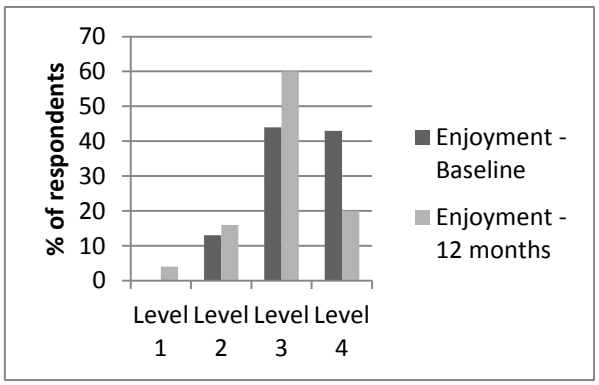
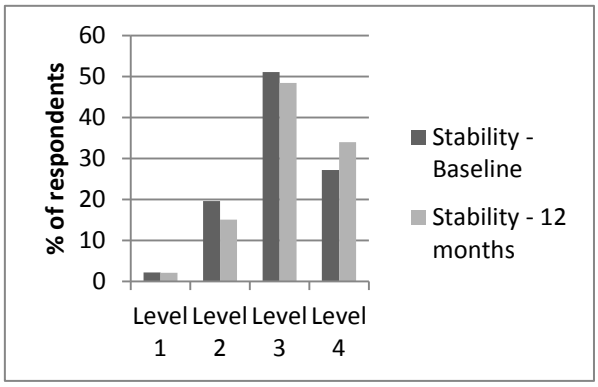


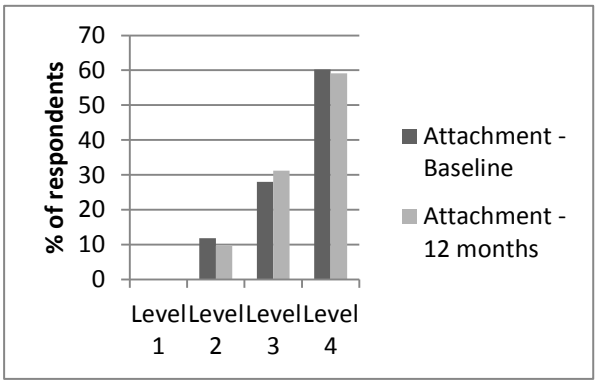
Figure B show the response profiles at baseline and follow-up the group of respondents reporting an improvement in their PQH-8 psychological health status (increase in PHQ-8 scores). These results are presented in numerical form in Appendix 29. Smaller changes in response profiles were found than in the group reporting a worsening of their PHQ-8 health status. The largest change was found in the Achievement item, with an increase of 14 percentage points of people answering level 4 (full capability). Smaller changes of between 1 and 7 percentage points were found in other items.

Figure B: ICECAP-O response profile at baseline and follow-up for participants reporting an improvement in their PHQ-8 health status

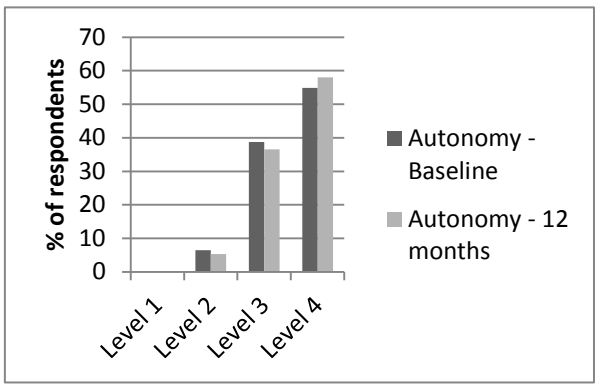
Stability Item



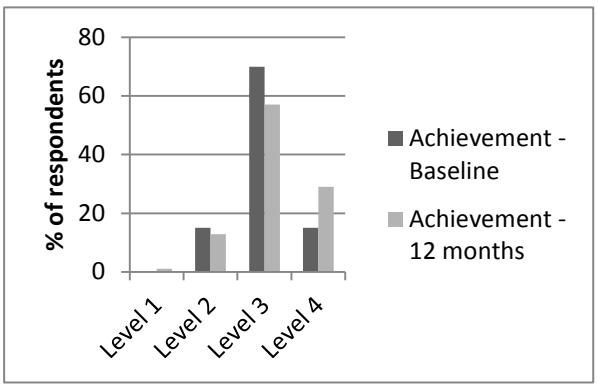
Attachment Item



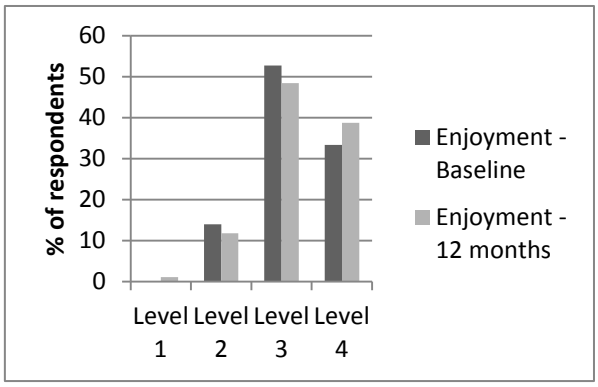
Autonomy Item



Achievement Item



Enjoyment Item



Non-weighted ICECAP-A score

Table B shows that cross-sectional correlations between the PHQ-8 scores and the non-weighted ICECAP-A scores at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of change in the scores of these measures between baseline and follow-up was weak and significant at the 1% level.

Table B: Cross-sectional and change correlations between the PHQ-8 and non-weighted ICECAP-A scores (n=331)

	ICECAP-A		
	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	
PHQ-8 score	-0.46**	-0.502**	-0.207**

* Significant at the 5% level, **Significant at the 1% level.

Table C shows change in non-weighted ICECAP-A scores by PHQ-8 anchor change groups. In the group of participants reporting an improvement in PHQ-8 health state the mean non-weighted ICECAP-A score increased. In the group reporting a reduction in their PHQ-8 health state, the mean non-weighted ICECAP-A score decreased. The change in ICECAP-A score was larger in the group that in had worsened than in the group that had improved. The effect sizes were very small for those who reported an improvement in PHQ-8 health status, while they were small for the group that worsened. For the ICECAP-A a larger change in scores was in the group that worsened than in the group that improved was found (Table C). For the EQ-5D-3L the larger change is in the group that improved rather than the group that worsened (Table D).

Table C: Mean change in non-weighted ICECAP-A scores by PHQ-8 anchor change groups (n=331)

Anchor group	Baseline ICECAP scores	12 month ICECAP scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	16.217	16.576	0.359 (-0.003, 0.720)	2.3%	0.15	0.2
No change	17.486	17.616	0.13 (-0.077, 0.336)	0.8%	0.07	0.09
Worsened	16.629	15.759	-0.87** (-1.398, -0.343)	5.8%	0.37	0.45

* Significant at the 5% level, **Significant at the 1% level.

Table D: Mean change in non-weighted EQ-5D-3L scores by GAD-7 anchor change groups (n=326)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	8.16	7.309	-0.852** (-1.142,-0.561)	8.5%	0.56	0.64
No change	7.421	7.015	-0.407** (-0.56,-0.253)	4.1%	0.35	0.36
Worsened	7.878	8.097	0.219 (-0.251, 0.69)	2.2%	0.14	0.15

* Significant at the 5% level, **Significant at the 1% level.

ICECAP-A tariff score

Table E shows that cross-sectional correlations between the PHQ-8 score and ICECAP-A tariff value at baseline and follow-up were moderate and significant at the 1% level. The correlation of change scores in these measures between baseline and follow-up was weak and significant at the 1% level.

Table E: Cross-sectional and change correlations between the PHQ-8 and ICECAP-A measures (n=331)

	ICECAP-A		
	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	
PHQ-8 score	-0.455**	-0.498**	0.190**

* Significant at the 5% level, **Significant at the 1% level.

Table F and Figure C show change in ICECAP-A tariff value by PHQ-8 anchor change groups. In the group of participants reporting an improvement in PHQ-8 scores the mean ICECAP-A tariff value increased. This increase was very small, non-significant and effect sizes were small. In participants who reported a decrease in their PHQ-8 scores the mean ICECAP-A tariff value decreased. This decrease was moderate, statistically significant at the 1% level and with moderate effects sizes. There are small reductions in the size of change as a percentage of possible change in the ICECAP-A tariff analysis in comparison to the non-weighted analysis.

Table F: Mean change in ICECAP-A tariff values by PHQ-8 anchor change groups (n=331)

Anchor group	Baseline ICECAP scores	12 month ICECAP scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.852	0.866	0.014 (-0.005, 0.032)	1.4%	0.11	0.15
No change	0.917	0.92	0.003 (-0.006, 0.011)	0.3%	0.02	0.03
Worsened	0.872	0.825	-0.048** (-0.078,-0.017)	4.8%	0.39	0.43

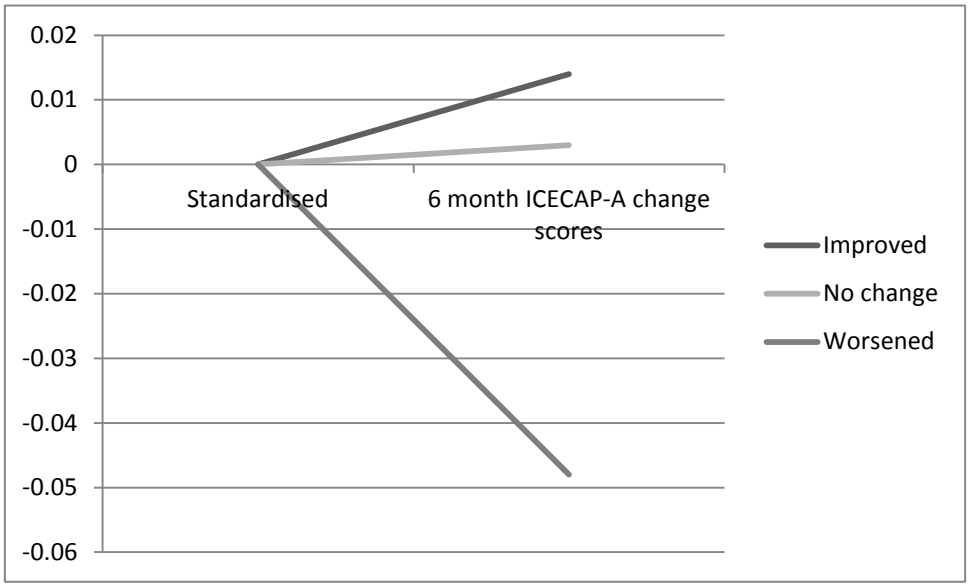
* Significant at the 5% level, **Significant at the 1% level.

Table G: Mean change in EQ-5D-3L index scores by GAD-7 anchor change groups (n=326)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.559	0.659	0.1** (0.05, 0.149)	6.3%	0.37	0.42
No change	0.689	0.744	0.056** (0.029,0.082)	3%	0.33	0.31
Worsened	0.653	0.621	-0.031 (-0.098,0.036)	2%	0.13	0.13

* Significant at the 5% level, **Significant at the 1% level.

Figure C: Mean change in ICECAP-O tariff values by PHQ-8 anchor change groups (n=331)



Appendix 28

ICECAP-A response profile for worsened PHQ-8 health status

	Baseline profile					12 month follow-up profile					Change between baseline and follow-up				
	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy
Level 1	2	0	0	0	0	4	2	0	2	4	+2	+2	0	+2	+4
Level 2	15	6	4	15	13	22	14	4	13	16	+7	+8	0	-2	+3
Level 3	53	35	35	59	44	56	33	33	67	60	+3	-2	-2	+8	+6
Level 4	30	59	61	26	43	18	51	64	18	20	-12	-8	+3	-8	-23

Appendix 29

ICECAP-A response profile for improved PHQ-8 health status

	Baseline profile					12 month follow-up profile					Change between baseline and follow-up				
	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy	Stab	Attach	Auto	Achieve	Enjoy
Level 1	2	0	0	0	0	2	0	0	1	1	0	0	0	+1	+1
Level 2	20	12	6	15	14	15	10	5	13	12	-5	-2	-1	-2	-2
Level 3	51	28	39	70	53	48	31	37	57	48	-3	+3	+2	+13	-5
Level 4	27	60	55	15	33	34	59	58	29	29	+7	-1	+3	+14	-4

Appendix 30

ICECAP-O response profile for worsened EQ-5D-3L index scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	2	8	3	3	0	2	12	3	5	3	0	+4	0	+2	+3
Level 2	12	20	15	10	10	12	18	16	18	12	0	-2	+1	+8	+2
Level 3	20	43	46	55	23	22	52	55	53	40	+2	+9	+9	+2	+17
Level 4	67	28	36	32	67	65	18	26	23	45	-2	-10	-10	-9	-22

Appendix 31

ICECAP-O response profile for improved EQ-5D-3L index scores.

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	1	7	2	2	2	0	2	3	1	1	-1	-5	+1	-1	-1
Level 2	7	27	22	19	9	8	28	15	19	8	+1	+1	-7	0	-1
Level 3	33	42	45	52	40	35	43	45	47	39	+2	+1	0	-5	-1
Level 4	59	25	30	27	48	57	27	37	37	52	-2	+2	+7	+10	+4

Appendix 32

EQ-5D-3L VAS anchor anlaysis

The anchor groups for the EQ-5D-3L VAS were formed based on minimally important difference values taken from Pickard, Neary and Cella[322] of 7 points. This study was completed in cancer patients, so limitations exist in applying this to a responsiveness assessment in the PastBP trial. The value from this study has been used because of the very limited number of studies that provided MID estimates for the EQ-5D-3L VAS. Groups were formed of participants who had improved or worsened by equal to or more than 7 points. EQ-5D-3L VAS change as a percentage of possible change was roughly 17% in both groups.

Table A: Group numbers and EQ-5D-3L VAS change scores in the EQ-5D-3L VAS anchor groups (n=273)

Anchor group	Number	Mean EQ-5D-3L VAS change in group (95% CI)	EQ-5D-3L VAS change as a % of possible change
Improved	94	17.37 (15.37, 19.36)	17.3%
No change	101	0.24 (-0.35, 0.82)	0.2%
Worsened	78	-17.06 (-18.79,-15.32)	17.1%

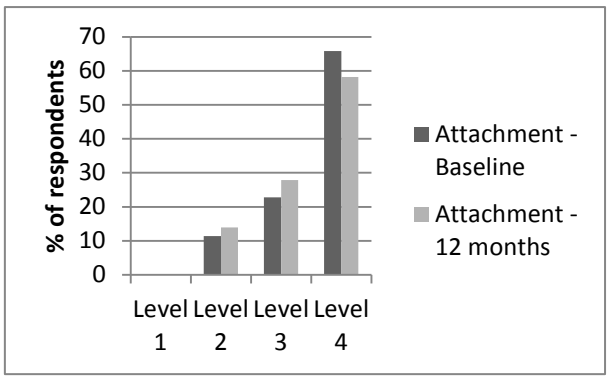
Item by item analysis

Figure A shows the ICECAP-O response profiles at baseline and follow-up for those who reported a worsening of their EQ-5D-3L VAS scores. The results are provided in numerical form in Appendix 33. Reductions of 8 points for the Attachment item, 5 points for the

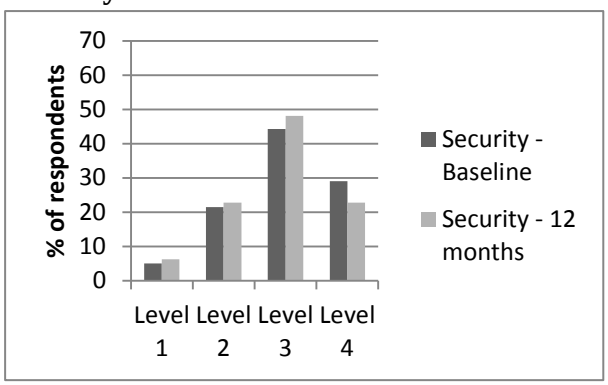
Security, Role and Enjoyment items and 2 points for the control item were found in the percentage of participants answering level 4 (full capability) for these items. Change in the percentage of respondents answering levels 1 and 2 were minimal in all items.

Figure A: ICECAP-O response profile at baseline and follow-up for participants reporting a reduction on their EQ-5D-3L VAS scores

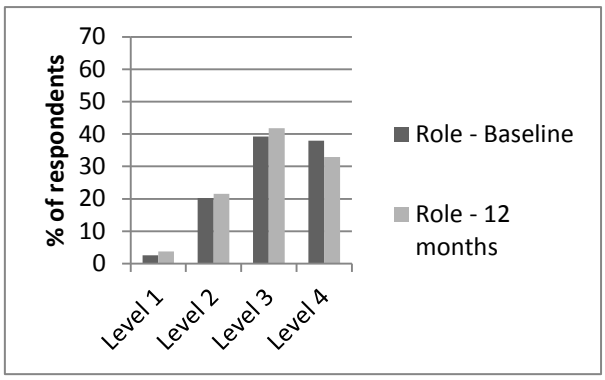
Attachment Item



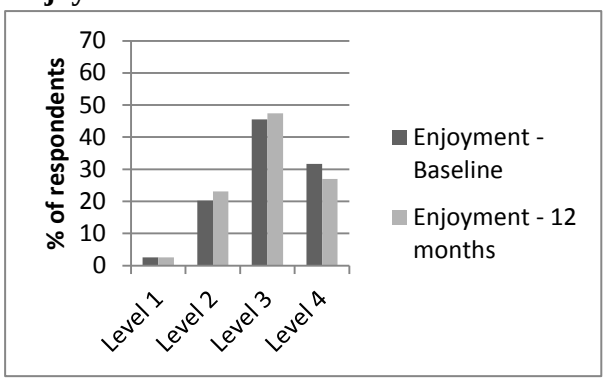
Security Item



Role Item



Enjoyment Item



Control Item

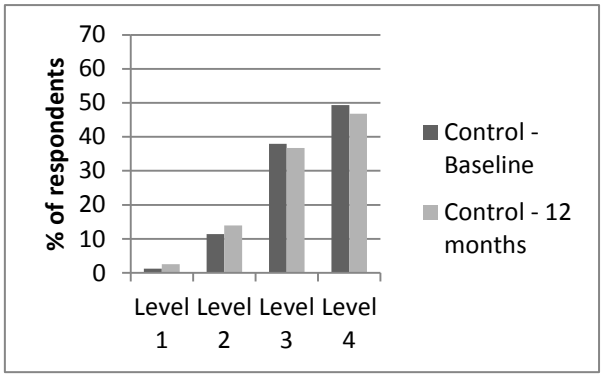
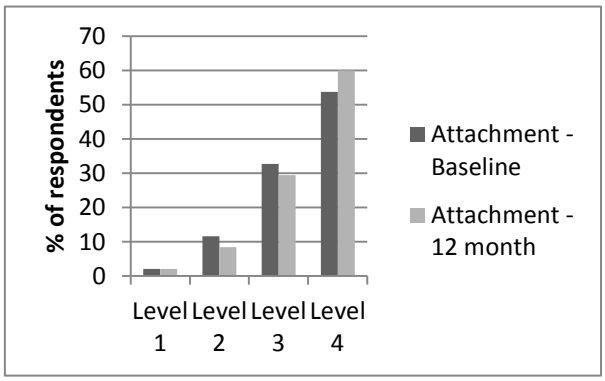


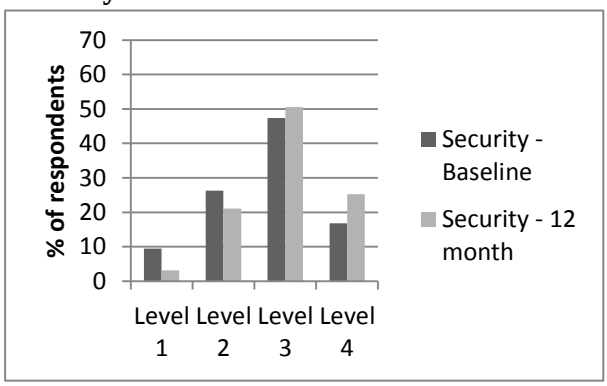
Figure B shows the ICECAP-O response profiles at baseline and follow-up for participants who reported an improvement in EQ-5D-3L VAS scores. The results are provided in numerical form in Appendix 34. The percentage of respondents who answered level 4 (full capability) for each item increased by 4 to 10 points between baseline and follow-up. The largest changes were in Security and Enjoyment, while the smallest change was in Control. There was a reduction in the percentage of participants answering in the bottom two levels for Attachment, Security, Role and Enjoyment, this was particularly marked for the Role and Security items.

Figure B: ICECAP-O response profile at baseline and follow-up for participants reporting an improvement on their EQ-5D-3L VAS scores

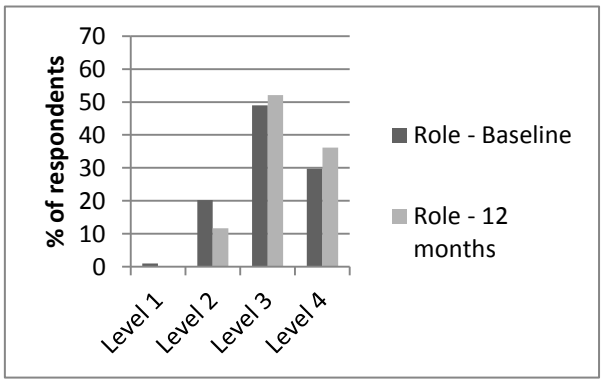
Attachment Item



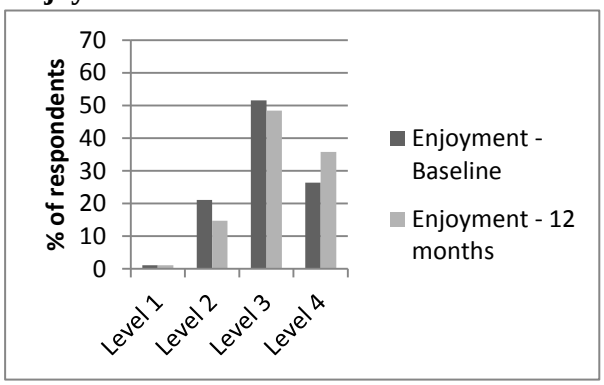
Security Item



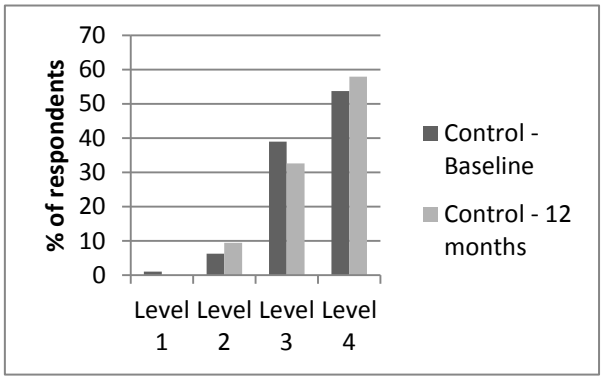
Role Item



Enjoyment Item



Control Item



Non-weighted ICECAP score analysis

The cross-sectional correlations between the EQ-5D-3L VAS and the non-weighted ICECAP-O scores at baseline and 12 month follow-up were moderate and significant at the 1% level. The correlation of the change in scores between baseline and 12 month follow-up in these two measures was moderate and significant at the 1% level. See Table B.

Table B: Cross-sectional and change correlations between EQ-5D-3L VAS and non-weighted ICECAP-O scores (n=273)

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	
EQ-5D-3L VAS	0.554**	0.493**	0.301**

* Significant at the 5% level, **Significant at the 1% level.

Table C shows change in non-weighted ICECAP-O score by EQ-5D-3L VAS anchor groups. In the groups of participants reporting an improvement in their EQ-5D-3L VAS scores the mean non-weighted ICECAP-O score value increased. In the group that reported a worsening of their EQ-5D-3L VAS scores the mean non-weighted ICECAP-O score decreased. The change in ICECAP scores in the group that had improved was greater than in the group that had worsened. The change in ICECAP-O scores as a percentage of possible change (Table C) was small in comparison to change as a percentage of possible change that occurred in the EQ-5D-3L VAS in each group (Table A). The effects sizes and SRMs for both change groups were small.

The use of non-weighted EQ-5D-3L score change (Table D) as a reference measure, allows a better understanding of the changes in the ICECAP -O. Change as a percentage of possible

change in the non-weighted EQ-5D scores and non-weighted ICECAP-O scores were similar: roughly 5% in the improved group and roughly 2.8% in the group that worsened. Effect sizes and SRMs for change in the EQ-5D-3L was similar to those for change in the ICECAP-O scores.

Table C: Mean change in non-weighted ICECAP-O scores by EQ-5D-3L VAS anchor change groups (n=273)

Anchor group	Baseline ICECAP scores	12 month ICECAP scores	Mean change in ICECAP tariff values (95% CI)	Change as % of possible change	ES	SRM
Improved	15.617	16.362	0.745** (0.341, 1.149)	4.97%	0.3	0.38
No change	16.564	16.564	0 (-0.357, 0.357)	0%	0.0	0.0
Worsened	16.063	15.645	-0.418 (-0.953, 0.117)	2.79%	0.14	0.17

* Significant at the 5% level, **Significant at the 1% level.

Table D: Mean change in EQ-5D-3L non-weighted scores by EQ-5D-3L VAS anchor change groups (n=286)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	6.745	7.245	0.5** (0.219,0.781)	5%	0.31	0.36
No change	6.619	6.628	0.009 (-0.224, -0.243)	0.1%	0.01	0.01
Worsened	7.494	7.217	0.277 (-0.573, 0.019)	2.8%	0.15	0.2

* Significant at the 5% level, **Significant at the 1% level.

ICECAP tariff analysis

Table E shows that the cross sectional correlations between the EQ-5D-3L VAS and ICECAP-O index scores at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of the change scores of these measures was weak and statistically significant at the 1% level.

Table E: Cross-sectional and change correlations between EQ-5D-3L VAS and ICECAP-O tariff value (n=273)

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	
EQ-5D-3L VAS	0.493**	0.458**	0.267**

* Significant at the 5% level, **Significant at the 1% level.

Table F and Figure C show change in ICECAP tariff value by EQ-5D-3L VAS anchor groups. In the group of participants reporting an improvement in EQ-5D-3L VAS scores the mean ICECAP-O tariff value increased. In the group reporting a worsening of EQ-5D-3L VAS scores the mean ICECAP-O tariff value decreased. Changes in the ICECAP-O tariff values were small and the change in ICECAP-O tariff values in the group reporting an improvement in EQ-5D-3L VAS score was statistically significant at the 1% level. In comparison to the non-weighted ICECAP analysis, change as a percentage of possible change was smaller in ICECAP-O tariff values than in the non-weighted ICECAP-O scores. This difference was particularly pronounced in the group that had improved; change as a percentage of possible change was 4.97% in the non-weighted analysis and 2.9% in the tariff analysis. Effects sizes and SRMs were small.

A comparison of the results of the ICECAP tariff analysis with the EQ-5D-3L index analysis (Table G) shows similarities. Effects sizes and SRMs were similar, while change as a percentage of possible change was 1 percentage point larger for the improved group in the EQ-5D-3L analysis.

Table F: Mean change in ICECAP-O tariff score by EQ-5D-3L VAS anchor change groups (n=273)

Anchor group	Baseline ICECAP scores	12 month ICECAP scores	Mean change in ICECAP tariff scores (95% CI)	Change as % of possible change	ES	SRM
Improved	0.843	0.872	0.029 ** (0.011, 0.048)	2.9%	0.22	0.32
No change	0.882	0.879	0.003 (-0.015, 0.022)	0.3%		
Worsened	0.857	0.837	-0.020 (-0.008, 0.049)	2.0%	0.16	0.16

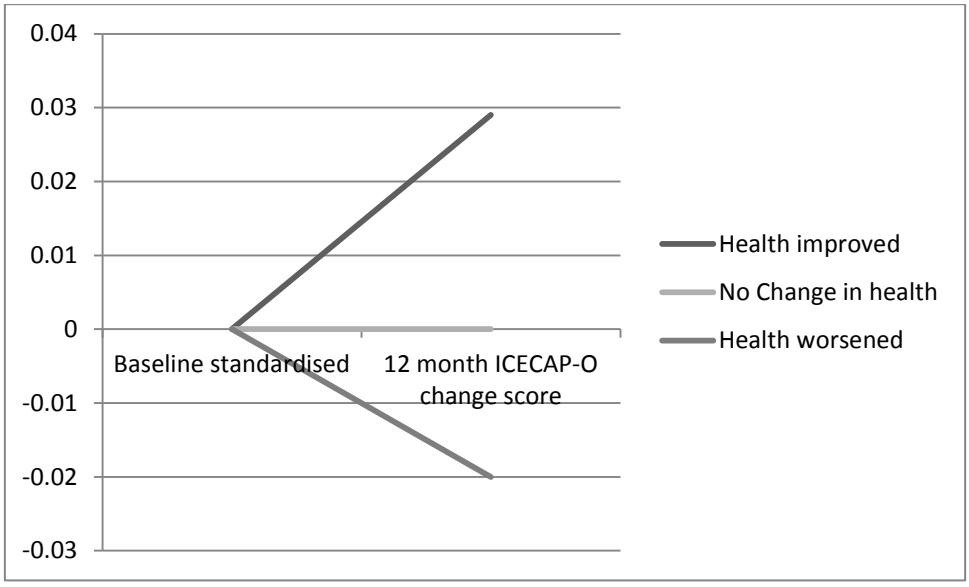
* Significant at the 5% level, **Significant at the 1% level.

Table G: Mean change in EQ-5D-3L index scores by EQ-5D-VAS anchor change groups (n=286)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean change in EQ-5D-3L index scores (95% CI)	Change as % of possible change	ES	SRM
Improved	0.717	0.779	0.062** (0.019,0.104)	3.9%	0.26	0.29
No change	0.781	0.796	0.015 (-0.018, 0.048)	0.9%	0.06	0.09
Worsened	0.718	0.681	-0.037 (-0.094, 0.02)	2.3%	0.14	0.14

* Significant at the 5% level, **Significant at the 1% level.

Figure C: Mean change in ICECAP-O scores by EQ-5D-3L VAS anchor change groups (n=273)



Appendix 33

ICECAP-O response profile for worsened EQ-5D VAS scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	5	3	2	1	0	6	4	2	2	0	+1	+1	0	+1
Level 2	11	21	20	20	11	14	23	21	23	14	+3	+2	+1	+3	+3
Level 3	23	44	39	46	38	28	48	42	47	37	+5	+4	+3	+1	-1
Level 4	66	29	38	32	49	58	23	33	27	47	-8	-5	-5	-5	-2

Appendix 34

ICECAP-O response profile for improved EQ-5D VAS scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	2	9	1	1	1	2	3	0	1	0	0	-6	-1	0	-1
Level 2	12	26	20	21	6	8	21	12	15	9	-4	-5	-8	-6	+3
Level 3	33	47	49	52	39	29	51	52	48	33	-4	+4	+3	-4	-6
Level 4	54	17	30	26	54	60	25	36	36	58	+6	+8	+6	+10	+4

Appendix 35

ICECAP-O response profile for improved Modified Rankin Scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	6	2	2	5	0	3	0	0	2	0	-3	-2	-2	-3
Level 2	13	22	22	16	6	13	22	9	16	5	0	0	-13	0	-1
Level 3	36	49	43	54	33	31	46	51	40	36	-5	-3	+8	-14	+3
Level 4	51	22	33	29	56	56	29	40	44	57	+5	+7	+7	+15	+1

Appendix 36

ICECAP-O response profile for worsened Modified Rankin Scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	6	0	0	0	0	6	4	1	3	0	0	+4	+1	+3
Level 2	10	23	14	21	9	15	20	19	17	11	+5	-3	+5	-4	+2
Level 3	28	42	49	47	33	25	51	47	55	38	-3	+9	+2	+8	+5
Level 4	62	29	37	32	58	60	22	30	27	49	-2	-7	-7	-5	-9

Appendix 37

SF-36 anchors anlaysis

This minimally important difference used to defined general health anchor groups was 7 [288]. Mean change in SF-36 general health sub-scale scores as a percentage of possible change was roughly 11% for both groups.

Table A: Numbers in groups and mean SF-36 general health sub-scale change scores in SF-36 general health sub-scale anchor (n=212)

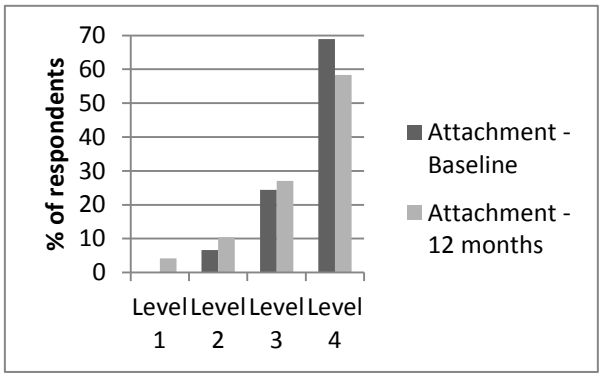
Anchor group	Number	Mean SF-36 general health change in group (95% CI)	Mean SF-36 general health change score as % of possible change
Improved	28	11.151 (9.832, 12.469)	11.2%
No change	146	-0.282 (-0.737, 0.173)	0.3%
Worsened	38	-10.819 (-12.202, -9.435)	10.8%

Item-by-item analysis

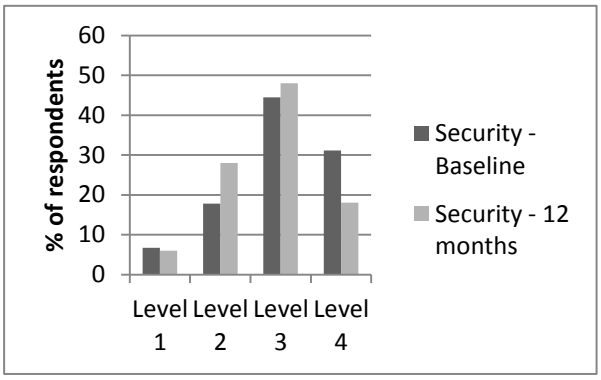
Figure A shows the ICECAP-O response profile at baseline and follow-up in respondents reporting a worsening of general health sub-scale scores. This analysis is presented in numerical form in Appendix 38. A reduction of 17 percentage points was seen in respondents answering level 4 (full capability) of Enjoyment. Decreases of between 8 and 13 percentage points were seen in respondents answering level 4 of Attachment, Security and Role. Minimal change was seen in the response profile of the Control item.

Figure A: ICECAP-O response profile at baseline and follow-up for participants reporting a worsening of their SF-36 general health sub-scale score.

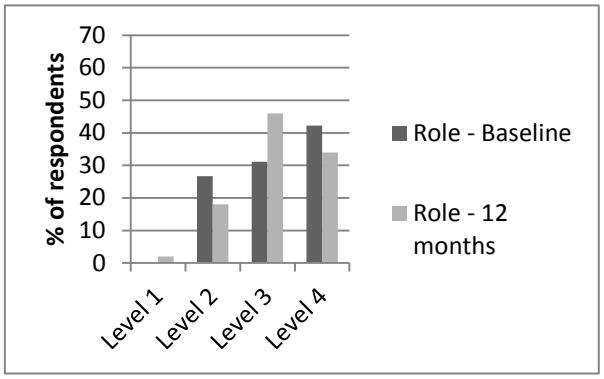
Attachment Item



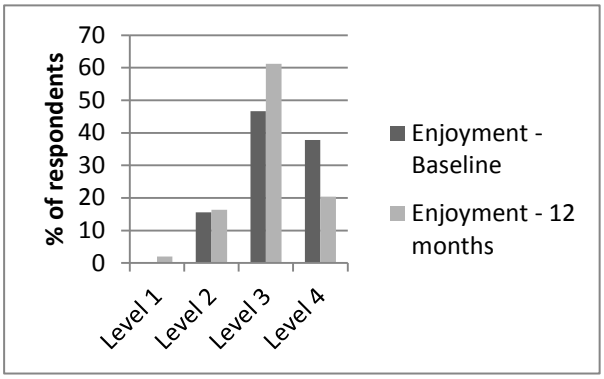
Security Item



Role Item



Enjoyment Item



Control Item

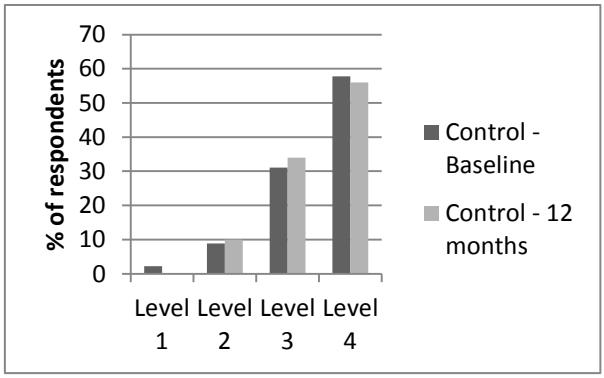
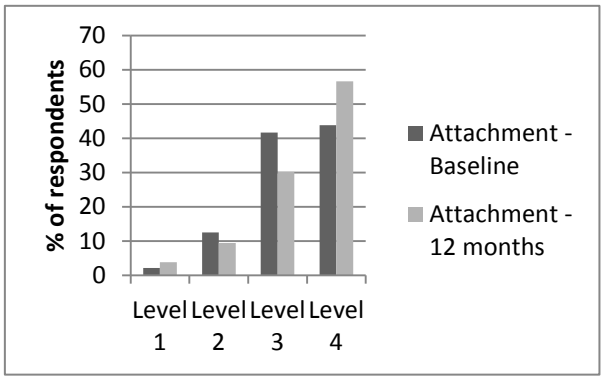


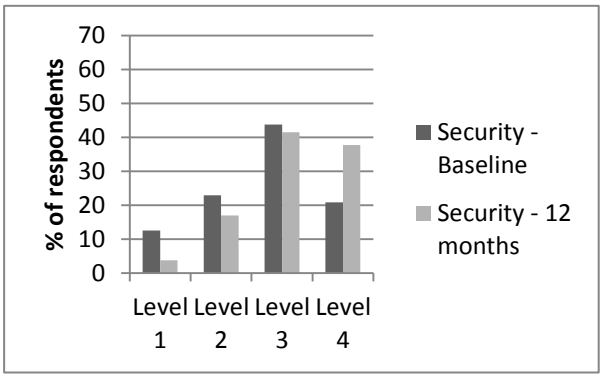
Figure B show the response profiles of the ICECAP-O at baseline and follow-up in the group of respondent reporting an improvement in general health sub-scale scores. This is presented in numerical form in Appendix 39. Minimal change was seen in response profile of the Control item. There was a 12 to 17 point increase between baseline and follow-up in the percentage of respondents answering level 4 (top level) of the other items. This change was most pronounced in the Security and Enjoyment items. There was a notable reduction in the percentage of respondents answering the bottom two levels of the Security and Role items.

Figure B: ICECAP-O response profile at baseline and follow-up for participants reporting an improvement in their SF-36 general health sub-scale score.

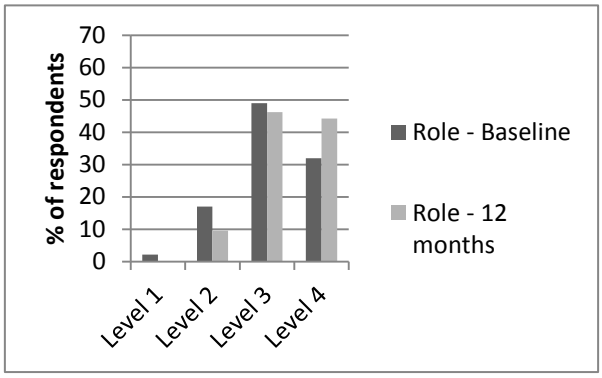
Attachment Item



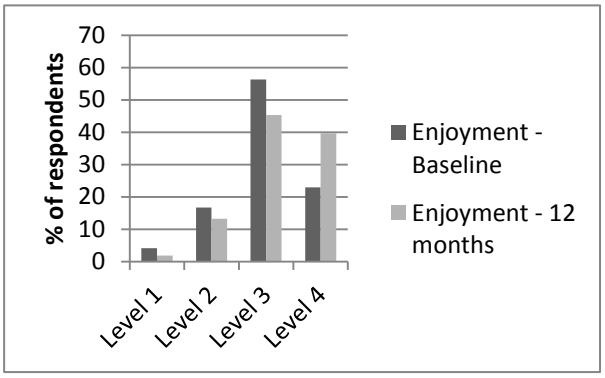
Security Item



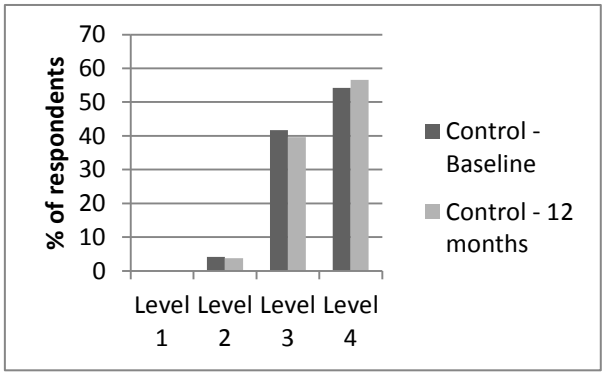
Role Item



Enjoyment Item



Control Item



Non-weighted ICECAP-O score analysis

Table B shows that the cross-sectional correlations between the SF-36 general health subscale and the non-weighted ICECAP-O score were moderate and significant at the 1% level. The correlation of the change in the two measures between baseline and follow-up was weak and significant at the 1% level.

Table B: Correlations between SF-36 general health scale and the ICECAP-O tariff value at baseline, 12 month follow-up and over time (n=212).

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	
SF-36 General Health	0.623**	0.526**	0.297**

* Significant at the 5% level, **Significant at the 1% level.

Table C shows change in non-weighted ICECAP-O score by SF-36 general health sub-scale anchor change groups. In the group of participants reporting an improvement in their SF-36 general health sub-scale score the mean non-weighted ICECAP-O score increased. This change was significant at the 1% level and medium (or approaching medium) effect size and SRM was found. Minimal change was seen in non-weighted ICECAP-O scores in participants reporting a worsening of their SF-36 sub-scale score.

A comparison of the changes in the non-weighted ICECAP-O analysis with the non-weighted EQ-5D-3L analysis (Table D) shows some differences. For the group reporting an improvement in general health scores, change as a percentage of possible change, effects sizes and SRMs were smaller for the EQ-5D-3L than for the IECCAP-O. In the group of respondents reporting a reduction in general health the reverse is true.

Table C: Mean change in non-weighted ICECAP-O scores by SF-36 general health sub-scale anchor change groups (n=212)

Anchor group	Baseline ICECAP scores	12 month ICECAP scores	Mean change (95% CI)	Change as % of possible change	ES	SRM
Improved	15.511	16.659	1.149** (0.513, 1.785)	7.6%	0.47	0.53
No change	16.371	16.530	0.158 (-0.129, 0.447)	0.1%	0.06	0.09
Worsened	16.419	16.209	-0.209 (-0.777, 0.359)	0.1%	0.08	0.11

* Significant at the 5% level, **Significant at the 1% level.

Table D: Mean change in EQ-5D-3L non-weighted scores by SF36 general health sub-scale anchor change groups (n=251)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	6.979	6.458	-0.521** (-0.894,-0.147)	5.2%	0.34	0.4
No change	6.924	6.645	-0.278 (-0.484,-0.073)	2.8%	0.17`	0.21
Worsened	7.2	7.578	0.378 (-0.003,0.758)	3.8%	0.26	0.29

* Significant at the 5% level, **Significant at the 1% level.

ICECAP-O tariff value analysis

The SF-36 general health sub-scale showed moderate, statistically significant correlation with the ICECAP-O tariff value at baseline and follow-up (Table E). The correlation in change scores between these measures was weak and statistically significant.

Table E: Correlations between SF-36 general health scale and the ICECAP-O tariff value at baseline, 12 month follow-up and over time (n=212).

	ICECAP-O		
	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	
SF-36 General Health	0.584**	0.523**	0.235**

* Significant at the 5% level, **Significant at the 1% level.

Table F and Figure C show change in ICECAP-O tariff value by SF-36 general health subscale anchor change groups. Participants who reported an increase in their general health score had a moderate mean change in their ICECAP-O tariff value of 0.042. This change in ICECAP-O tariff value between baseline and follow-up was significant at the 1% level and the effect size and SRM was small. In the group of participants reporting a decrease in their general health scale score minor, non-statistically significant change in ICECAP-O tariff values was seen.

Differences exist between the SF-36 general health sub-scale non-weighted ICECAP-O analysis and ICECAP-O tariff analysis. In the group of respondents who reported an improvement in general health a substantial difference in change as a percentage of possible change was found between the non-weighted (7.6%) and tariff analysis (4.2%). The effect sizes and SRM for the tariff value was also smaller than for the non-weighted score. In the group of respondents who reported a worsening of general health, change as a percentage of possible change was slightly larger in the tariff analysis in comparison to the non-weighted analysis.

Table F: Mean change in ICECAP-O tariff value by SF-36 general health sub-scale anchor change groups (n=212)

Anchor group	Baseline scores	12 month follow-up scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.842	0.884	0.042** (0.013, 0.07)	4.2%	0.31	0.43
No change	0.872	0.878	0.006 (0.019, -0.008)	0.4%	0.01	0.01
Worsened	0.871	0.86	-0.008 (0.024, 0.04)	0.8%	0.07	0.08

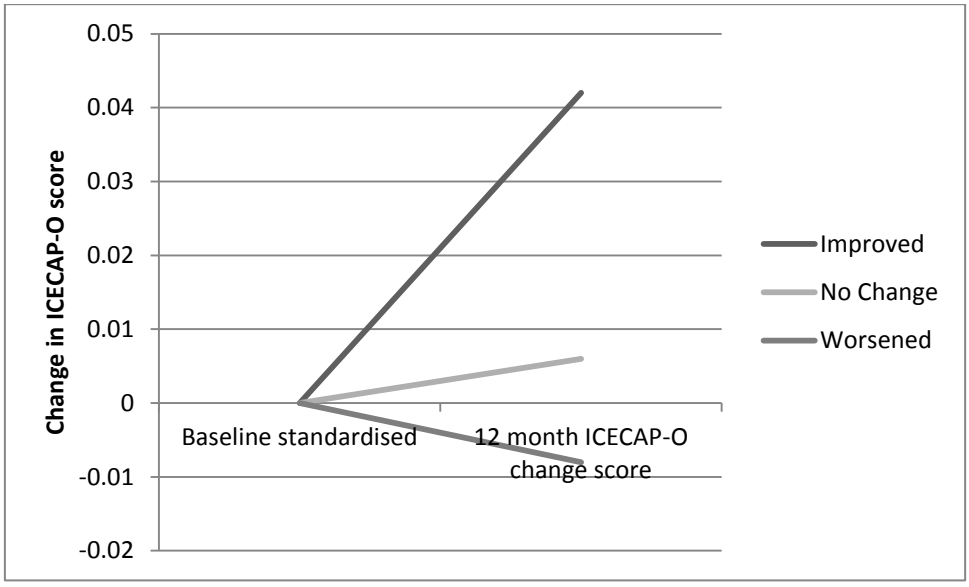
* Significant at the 5% level, **Significant at the 1% level.

Table G: Mean change in EQ-5D-3L index scores by SF36 general health sub-scale anchor change groups (n=251)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.76	0.80	0.049 (-0.004,0.102)	3.1%	0.23	0.27
No change	0.751	0.799	0.048** (0.015, 0.081)	3%	0.21	0.23
Worsened	0.736	0.693	-0.043 (0.096,-0.009)	2.7%	0.2	0.24

* Significant at the 5% level, **Significant at the 1% level.

Figure C: Mean change in ICECAP-O tariff value by SF-36 general health sub-scale anchor change groups (n=212)



Appendix 38

ICECAP-O response profile for worsened SF-36 general health sub-scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	7	0	0	2	4	6	2	2	0	+4	-1	+2	+2	-2
Level 2	7	18	27	15	9	10	28	18	16	10	+3	+10	-9	+1	+1
Level 3	24	44	31	47	31	27	48	46	61	34	+3	+4	+15	+14	+3
Level 4	69	31	42	38	58	58	18	34	20	56	-11	-13	-8	-18	-2

Appendix 39

ICECAP-O response profile for improved SF-36 general health sub-scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	2	12	2	4	0	4	4	0	2	0	+4	-8	-2	-2	0
Level 2	12	23	17	17	4	9	17	10	13	4	-3	-6	-7	-4	0
Level 3	42	44	49	56	42	30	41	46	45	40	-12	-3	-3	-11	-2
Level 4	44	21	32	23	54	57	38	44	40	57	+13	+17	+12	+17	+3

Appendix 40

SF-36 vitality sub-scale anchor analysis

The minimally important difference value used to form the vitality anchor groups was 6.7 [288]. The mean change in SF-36 vitality sub-scale scores was roughly 11 in both the improved and worsened groups.

Table A: Numbers and the mean change in SF-36 sub-scale vitality scores in each anchor group.

Anchor group	Number	Mean SF-36 vitality change in group (95% CI)
Improved	41	11.238 (9.757, 12.719)
No change	154	-0.157 (-0.662, 0.348)
Worsened	38	-10.641 (-11.846, -9.4360

Figure A shows the response profiles at baseline and follow-up in the group of respondents who reported worsening in vitality sub-scale scores. These are reported in numerical form in Appendix 41. The response profiles changed. Changes of between 6 and 9 points were seen in the percentage of respondents reporting full capability for Security, Role and Enjoyment. Minimal change was found in the percentage of people reporting full capability for Attachment and Control.

Item-by-item analysis

Figure A: ICECAP-O response profiles (percentages of respondents answering each level of each item) at baseline and follow-up for participants reporting a worsening of their SF-36 vitality sub-scale score.

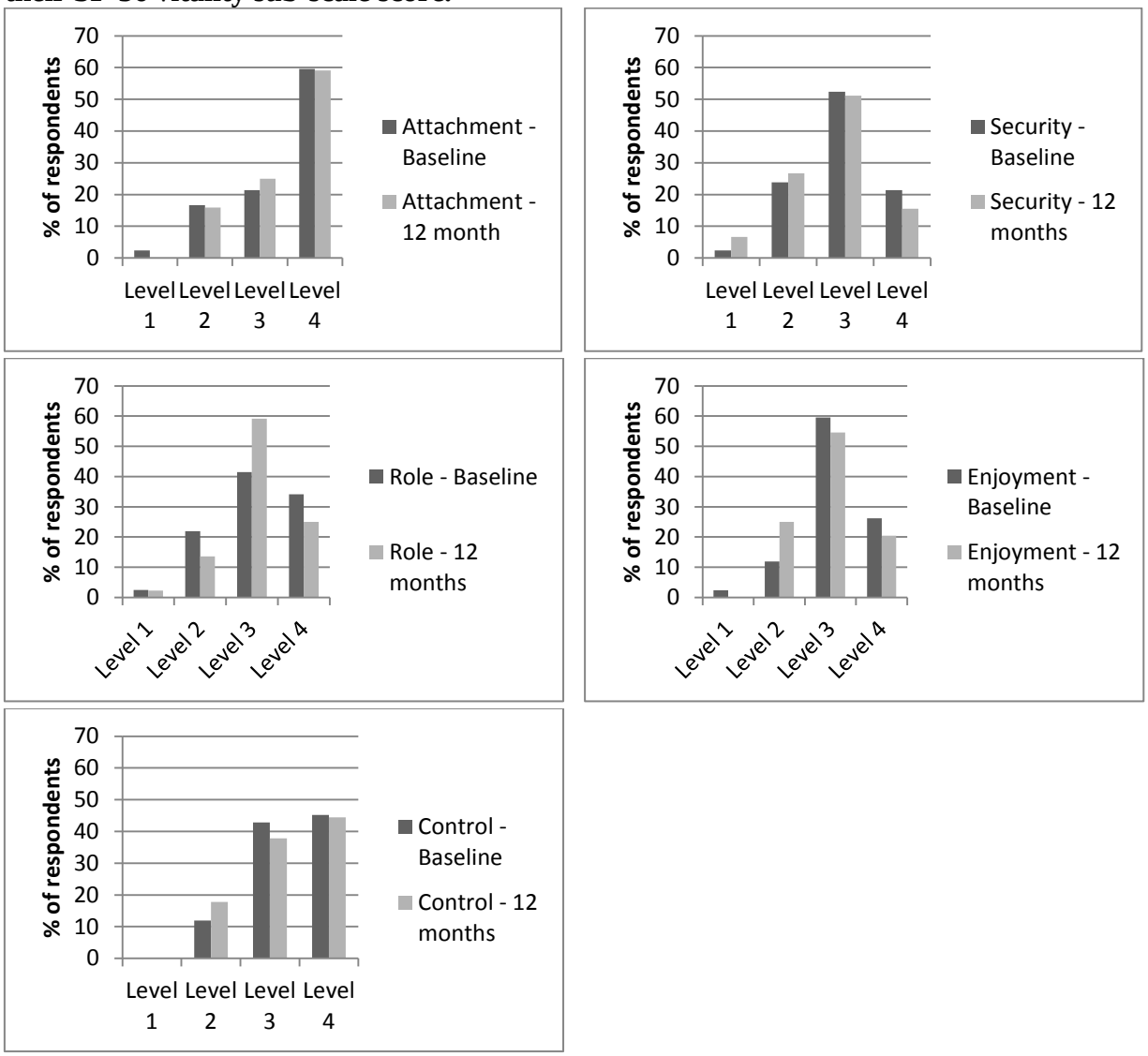
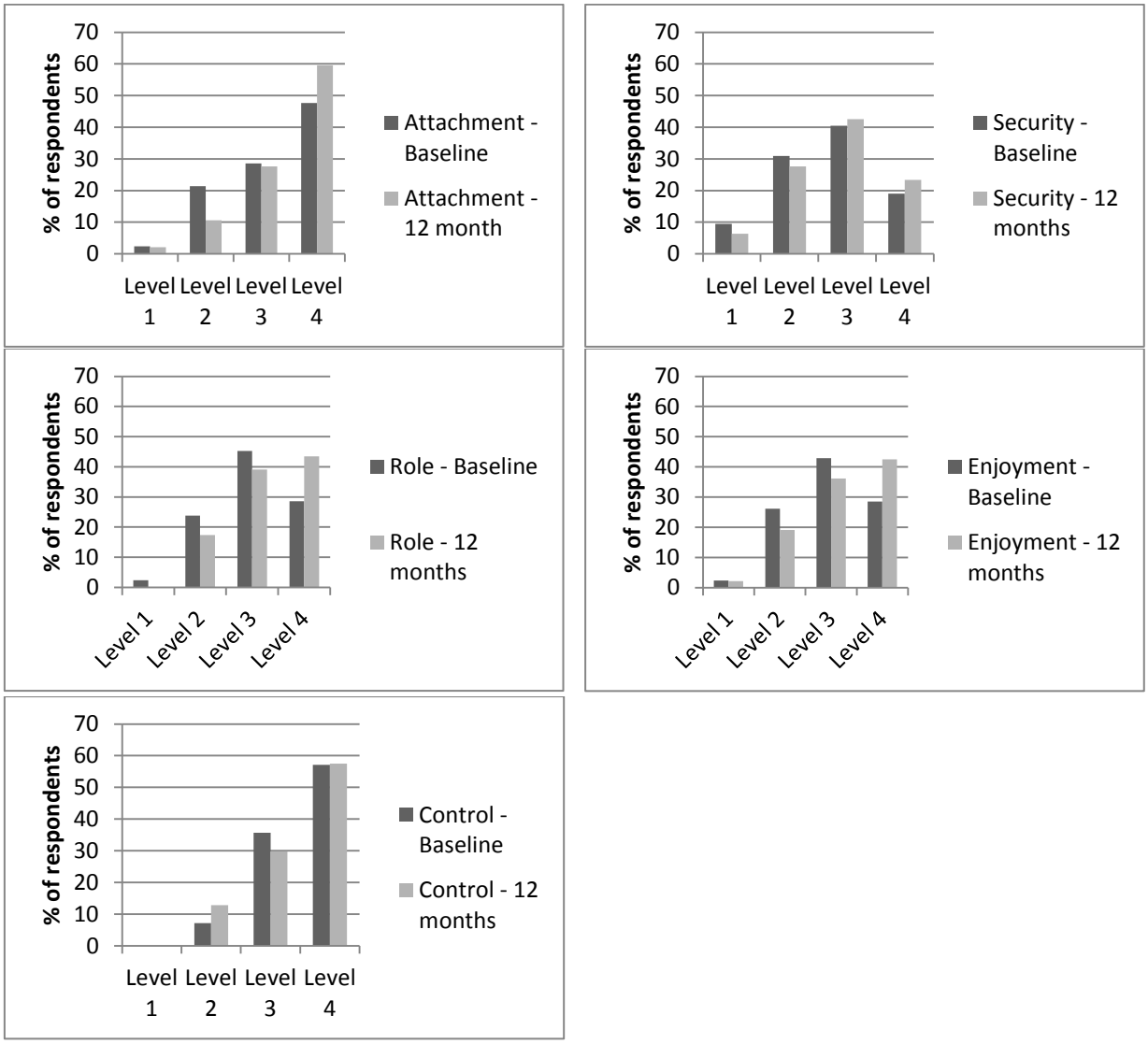


Figure B shows the response profiles at baseline and follow-up for the group of respondents who reported and improvement in their vitality sub-scale scores. These are reported in numerical form in Appendix 42. Changes of 14 points were seen for Role and Enjoyment,

and a change of 11 points for Attachment, in the percentage of respondents reporting full capability. Smaller changes were seen for Security and Control

Figure B: Baseline and 12 month follow-up response profiles for each item of the ICECAP-O for participants reporting a improvement in their SF-36 vitality sub-scale score.



Non-weighted ICECAP-O score analysis

Table B shows the cross-sectional correlations between SF-36 vitality sub-scale score and the ICECAP-O score at baseline and follow-up were moderate and statistically significant at the

1% level. The correlation of change in scores between baseline and follow-up in these measures was weak and significant at the 1% level.

Table B: Correlations between SF-36 vitality scale and the ICECAP-O tariff score at baseline, 12 month follow-up and over time (n=233).

	Cross sectional correlation		Change correlation
	Baseline	12 month follow-up	Baseline to 12 month follow up
SF-36 Vitality	0.56**	0.55**	0.19**

** Difference significant at p<0.01

Table C shows change in non-weighted ICECAP-O score by SF-36 vitality sub-scale anchor groups. In the group reporting an improvement in vitality scores the mean non-weighted ICECAP-O score increased. In the group reporting a worsening of vitality scores the mean ICECAP-O scores decreased. The change in ICECAP-O scores in the group that improved was larger than in the group that had worsened. The effect sizes and SRM for the improved group were small and moderate respectively, while for the worsened group they were very small.

The use of the non-weighted EQ-5D-3L score (Table D) as a comparator, allows a better understanding of the changes in the ICECAP-O measure. In the group that reported improved vitality scores change as a percentage of possible change was larger for the ICECAP-O than for the EQ-5D-3L. The reverse is true for the group that worsened.

Table C: Mean change in ICECAP-O score by SF-36 vitality sub-scale change (n=233)

Anchor group	Baseline scores	12 month follow-up scores	Mean change (95% CI)	Change as % of possible change	ES	SRM
Improved	15.366	16.585	1.219** (0.631, 1.807)	8.1%	0.43	0.65
No change	16.361	16.348	-0.013 (-0.339, 0.313)	0.1%	0.01	0.01
Worsened	15.718	15.410	-0.308 (-0.912, 0.297)	2.0%	0.11	0.16

** Difference significant at p<0.01

Table D: Mean change in EQ-5D-3L non-weighted scores by SF36 vitality health sub-scale change (for comparison)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	7.208	6.583	-0.625 (-1.059, -0.19)	6.2%	0.36	0.41
No change	6.894	6.77	-0.124 (-0.328, 0.08)	1.2%	0.07	0.09
Worsened	7.357	7.881	0.524 (0.121,0.926)	5.2%	0.28	0.4

ICECAP-O tariff score analysis

Table E shows the cross-sectional correlations between SF-36 vitality sub-scale score and the ICECAP-O value tariff at baseline and follow-up were moderate and statistically significant at the 1% level. The correlation of change in scores between baseline and follow-up in these measures was weak and significant at the 1% level.

Table E: Correlations between SF-36 vitality scale and the ICECAP-O tariff score at baseline, 12 month follow-up and over time (n=233).

	Cross sectional correlation		**Change correlation**
	Baseline	12 month follow-up	Baseline to 12 month follow up
SF-36 Vitality	0.56**	0.55**	0.19**

** Difference significant at p<0.01

Table F and Figure C shows change in ICECAP-O value tariff by SF-36 vitality sub-scale anchor groups. In the group reporting an improvement in vitality scores the mean ICECAP-O value tariff increased. In the group reporting a worsening of vitality scores there was minimal change in the mean ICECAP-O value tariff. The effect sizes and SRM for the improved group were small and moderate respectively. In comparison to the non-weighted ICECAP-O score the change as a percentage of possible change was smaller in the ICECAP-O value tariff analysis.

Table F: Mean change in ICECAP-O tariff score by SF-36 vitality sub-scale change (n=233)

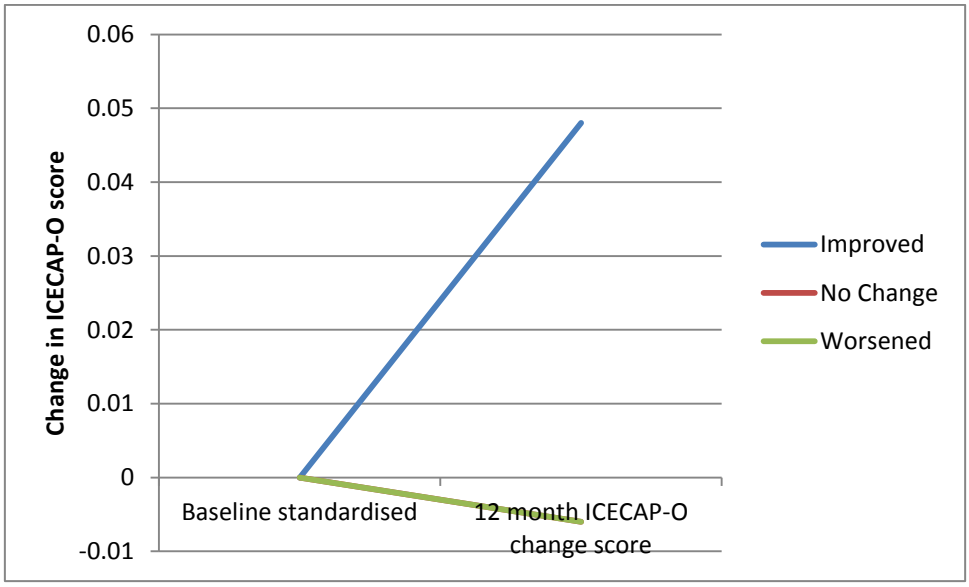
Anchor group	Baseline scores	12 month follow-up scores	Mean ICECAP-O change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.825	0.873	0.048** (0.074, 0.023)	4.8%	0.32	0.6
No change	0.874	0.868	-0.006 (0.011, -0.023)	0.6%		
Worsened	0.838	0.832	-0.006 (0.031, -0.043)	0.6%	0.04	0.07

** Difference significant at p<0.01

Table G: Mean change in EQ-5D-3L non-weighted scores by SF36 vitality health sub-scale change (for comparison)

Anchor group	Baseline EQ-5D-3L scores	12 month EQ-5D-3L scores	Mean EQ-5D-3L change (95% CI)	Change as % of possible change	ES	SRM
Improved	0.715	0.798	0.084* (0.013,0.154)	5.3%	0.35	0.34
No change	0.751	0.774	0.022 (0.009,0.054)	1.4%	0.09	0.11
Worsened	0.712	0.642	-0.069 (-0.146,0.007)	4.3%	0.35	0.28

Figure C: mean change in ICECAP-O tariff score by SF-36 vitality sub-scale change



Appendix 41

ICECAP-O response profile for worsened SF-36 vitality sub-scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	2	2	2	2	0	0	7	2	0	0	-2	+5	0	-2	0
Level 2	17	24	22	12	12	16	27	14	25	18	-1	+3	-8	+13	+6
Level 3	21	52	42	60	43	25	51	59	55	38	+4	-1	+17	-5	-5
Level 4	60	21	34	26	45	59	15	25	20	44	-1	-6	-9	-6	-1

Appendix 42

ICECAP-O response profile for improved SF-36 vitality sub-scale scores

	Baseline profile					**12 month follow-up profile**									
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	2	10	2	2	0	2	6	0	2	0	0	-4	-2	0	0
Level 2	21	31	24	26	7	11	28	17	19	13	-10	-3	-8	-7	+6
Level 3	29	40	45	43	36	28	43	39	36	30	-1	+3	-6	-7	-6
Level 4	48	19	29	29	57	59	23	43	43	57	+11	+4	+14	+14	0

Appendix 43

ICECAP-O response profile for worsened SF-36 social function sub-scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	2	4	0	0	2	2	10	2	3	2	0	+6	+2	+3	0
Level 2	8	30	23	23	11	12	24	15	22	12	+4	-6	-8	-1	+1
Level 3	30	45	43	47	34	32	44	54	56	47	+2	-1	+11	+9	+13
Level 4	60	21	34	29	53	54	22	29	19	39	-6	+1	-5	-10	-14

Appendix 44

ICECAP-O response profile for improved SF-36 social function sub-scale scores

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	12	5	5	2	0	4	0	0	0	0	-8	-5	-5	-2
Level 2	24	29	27	21	10	16	24	18	20	16	-8	-5	-9	-1	+6
Level 3	26	52	51	55	45	40	51	55	44	40	+14	-1	+4	-11	-5
Level 4	50	7	17	19	43	44	20	27	36	44	-6	+13	+10	+17	+1

Appendix 45

Change in individual symptoms for improved SSE

Symptom	Percentage at baseline		Percentage at follow-up	Percentage change
Pain	76.7%		39.5%	37.2%
Sore throat	27.9%		4.6%	23.3%
Nausea	11.6%		9.3%	2.3%
Breathlessness	48.8%		18.6%	30.2%
Weight loss	11.6%		11.6%	0%
Fatigue	69.8%		39.5%	30.3%
Stiff joints	72.1%		51.1%	21%
Sore eyes	16.3%		16.3%	0%
Wheeziness	23.3%		11.6%	11.7%
Headaches	34.9%		20.9%	14%
Upset stomach	25.6%		11.6%	14%
Sleep difficulties	53.5%		20.9%	32.6%
Dizziness	34.9%		11.6%	23.3%
Loss of strength	34.9%		25.6%	9.3%
Los of libido	27.9%		13.9%	14%
Impotence	25.6%		13.9%	11.7%
Feeling flushed	18.6%		16.3%	2.3%
Fast heart rate	34.9%		9.3%	25.6%
Pins and needles	32.6%		23.6%	9%
Cough	41.9%		23.3%	18.6%
Swelling of legs/ankles	25.6%		23.3%	2.3%
Mood change	18.6%		7%	11.6%
Rash	9.3%		4.6%	4.7%
Dry mouth	27.9%	9.3%	18.6%	

Change in individual symptoms for worsened SSE

Symptom	Percentage at baseline	Percentage at follow-up	Percentage change
Pain	37.5%	81.2%	43.7%
Sore throat	3.1%	25%	21.9%
Nausea	9.3%	15.6%	6.3%
Breathlessness	31.2%	59.4%	28.2%
Weight loss	3.1%	18.7%	15.6%
Fatigue	31.2%	68.7%	37.5%
Stiff joints	43.7%	59.4%	15.7%
Sore eyes	9.4%	28.1%	18.7%
Wheeziness	15.6%	31.2%	15.6%
Headaches	9.4%	21.9%	12.5%
Upset stomach	9.4%	18.7%	9.3%
Sleep difficulties	31.2%	50%	18.8%
Dizziness	18.7%	40.6%	21.9%
Loss of strength	12.5%	40.6%	28.1%
Los of libido	18.7%	31.2%	12.5%
Impotence	18.7%	25%	6.3%
Feeling flushed	12.5%	21.9%	9.4%
Fast heart rate	15.6%	25%	9.4%
Pins and needles	31.2%	31.2%	0%
Cough	31.2%	43.7%	12.5%
Swelling of legs/ankles	25%	53.1%	28.1%
Mood change	6.2%	25%	18.8%
Rash	6.2%	18.7%	12.5%
Dry mouth	12.5%	40.6%	28.1%

Appendix 46

ICECAP-O response profiles for increase in symptoms and side-effects

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	4	0	0	0	0	4	4	4	4	0	0	+4	+4	+4
Level 2	7	18	7	15	4	11	26	11	11	4	+4	+8	+4	-4	0
Level 3	18	48	37	44	37	30	44	48	54	37	+12	-4	+11	+10	0
Level 4	74	30	56	41	59	59	26	37	31	55	-15	-4	-19	-10	-4

Appendix 47

ICECAP-O response profiles for reduction in symptoms and side-effects

	Baseline profile					**12 month follow-up profile**					**Change between baseline and follow-up**				
	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont	Attach	Sec	Role	Enjoy	Cont
Level 1	0	8	0	0	3	0	5	0	3	3	0	-3	0	+3	0
Level 2	8	17	26	22	17	3	14	23	11	11	-5	-3	-3	-11	-6
Level 3	22	44	40	47	28	33	42	26	42	47	+11	-2	-14	-5	+19
Level 4	69	31	34	31	53	64	39	51	44	39	-5	+8	+15	+13	-14

Reference List

1. Robeyns I (2005) The capability approach: a theoretical survey. Journal of Human Development and Capabilities 6: 93-117.

2. Hicks JR (1939) The Foundations of Welfare Economics. Economic Journal 49: 696-712.

3. Culyer AJ (1991) The normative economics of health care finance and provision. In: McGuire A, editors. Providing health care: The economics of alternative systems of finance and delivery. Oxford: Oxford University Press.

4. Coast J, Smith RD, Lorgelly P (2008) Welfarism, extra-welfarism and capability: the spread of ideas in health economics. Soc Sci Med 67: 1190-1198.

5. Keeley T (2009) A welfarist approach is the best perspective for making resource allocation decisions in health care. Discuss.

6. Boadway, R. W. and Bruce, N. (1984) Welfare Economics. Oxford: Basil Blackwell Publisher Limited.

7. Brouwer WB, Culyer AJ, van Exel NJ, Rutten FF (2008) Welfarism vs. extra-welfarism. J Health Econ 27: 325-338.

8. Hurley J (2000) An overview of the normative economics of the health care sector. In: Culyer AJ, Newhouse J, editors. Handbook of Health Economics. Amsterdam: Elsevier Science. pp. 55-118.

9. Brouwer WB, Culyer AJ, van Exel NJ, Rutten FF (2008) Welfarism vs. extra-welfarism. J Health Econ 27: 325-338.

10. Boadway, R. W. and Bruce, N. (1984) Welfare Economics. Oxford: Basil Blackwell Publisher Limited.

11. Pigou, A. C. (1920) The Economics of Welfare. London: Macmillan.

12. Robbins, L. (1932) An Essay on the nature and significance of economic science. London: Macmillan.

13. Suzumura K (2000) Welfare economics beyond welfarist-consequentialism. Japanese Economic Review 51: 1-32.

14. van Praag BMS (1993) The relativity of the welfare concept. In: Nussbaum MC, Sen AK, editors. The Quality of Life. Oxford: Clarendon Press.

15. Bentham, J. (1789) An introduction to the principle of morals and legislation. Oxford: Blackwell.

16. Dolan P, Kahneman D (2008) Interpretations of utility and their implications for the valuation of health. Economic Journal 118: 215-234.

17. Kahneman D, Wakker PP, Sarin R (1997) Back to Bentham? - Explorations of experienced utility. Quarterly Journal of Economics 112: 375-405.

18. Edgeworth, F. (1881) Mathematical psychics. New York: Kelley.

19. Kahneman D, Sugden R (2005) Experienced utility as a standard of policy evaluation. Environmental & Resource Economics 32: 161-181. DOI 10.1007/s10640-005-6032-4.

20. Kahneman D, Wakker PP, Sarin R (1997) Back to Bentham? - Explorations of experienced utility. Quarterly Journal of Economics 112: 375-405.

21. Pareto, V. (1909) Manuel d'economii politique. Paris.

22. Hicks JR (1939) The Foundations of Welfare Economics. Economic Journal 49: 696-712.

23. Fisher I (1918) Is "Utility" the Most Suitable Term for the Concept It Is Used to Denote? American Economic Review 8: 335-337.

24. Dolan P, Kahneman D (2008) Interpretations of utility and their implications for the valuation of health. Economic Journal 118: 215-234.

25. Sieff EM, Dawes RM, Loewenstein G (1999) Anticipated versus actual reaction to HIV test results. American Journal of Psychology 112: 297-311.

26. Redelmeier DA, Kahneman D (1996) Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. Pain 66: 3-8.

27. Mooney, G. H. (1994) Key Issues in Health Economics. London: Harvester Wheatsheaf.

28. Birch S, Melnikow J, Kuppermann M (2003) Conservative versus aggressive follow up of mildly abnormal Pap smears: Testing for process utility. Health Economics 12: 879-884.

29. Benz M, Stutzer A (2003) Do workers enjoy procedural utility. Applied Economics Quarterly 49: 149-172.

30. Hahn F (1982) On some difficulties of the utilitarian economist. In: Sen AK, Williams B, editors. Utilitarianism and Beyond. Cambridge: Cambridge University Press.

31. Frey BS, Benz M, Stutzer A (2004) Introducing procedural utility: Not only what, but also how matters. Journal of Institutional and Theoretical Economics-Zeitschrift fur Die Gesamte Staatswissenschaft 160: 377-401.

32. Tsuchiya A, Miguel LS, Edlin R, Wailoo A, Dolan P (2005) Procedural justice in public health care resource allocation. Appl Health Econ Health Policy 4: 119-127. 426 [pii].

33. Culyer AJ (1991) The normative economics of health care finance and provision. In: McGuire A, editors. Providing health care: The economics of alternative systems of finance and delivery. Oxford: Oxford University Press.

34. Davis JB, McMaster R (2007) The individual in mainstream health economics: A case of Persona non-grata. Health Care Analysis 15: 195-210.

35. Morris, S., Devlin, N., and Parkin, D. (2007) Economic anaysis in health care. Chichester: John Wiley and Sons.

36. Ng, Y. (1979) Welfare Economics: Introduction and development of basic concepts. London: The MacMillan Press Ltd.

37. Kaldor N (1939) Welfare Propositions of Economics and Interpersonal Comparisons of Utility. Economic Journal 49: 549-552.

38. Kaldor N (1939) Welfare Propositions of Economics and Interpersonal Comparisons of Utility. Economic Journal 49: 549-552.

39. Robinson R (1993) Cost-benefit analysis. BMJ 307: 924-926.

40. Sugden R, Williams A (1978) The Objective in Cost-benefit Analysis. In: The Principles of Practical Cost-Benefit Analysis. Oxford: Oxford University Press. pp. 89-98.

41. Diener A, O'Brien B, Gafni A (1998) Health care contingent valuation studies: a review and classification of the literature. Health Econ 7: 313-326.

42. McIntosh E, Donaldson C, Ryan M (1999) Recent advances in the methods of cost-benefit analysis in healthcare. Matching the art to the science. Pharmacoeconomics 15: 357-367.

43. Weisbrod BA (1983) A guide to benefit-cost analysis, as seen through a controlled experiment in treating the mentally ill. J Health Polit Policy Law 7: 808-845.

44. Olsen JA, Donaldson C (1998) Helicopters, hearts and hips: using willingness to pay to set priorities for public sector health care programmes. Soc Sci Med 46: 1-12.

45. Stevens TH, Belkner R, Dennis D, Kittredge D, Willis C (2000) Comparison of contingent valuation and conjoint analysis in ecosystem management. Ecological Economics 32: 63-74.

46. Hjalte K, Norinder A, Trawen A (2010) Conjoint Analysis versus Contigent Valuation: estimating Risk Values and Death Risk Equivalents in Road Traffic. European Transport Conference 2000 139-147.

47. Arrow K, Solow R, Portney PR, Leamer EE, Radner R, Schuman H (1993) Report of the NOAA Panel on Contingent Valuation. 1-66.

48. Sen, A. (1987) On Ethics and Economics. Oxford: Basil Blackwell Ltd.

49. Coast J (2009) Maximisation in extra-welfarism: A critique of the current position in health economics. Soc Sci Med 69: 786-792.

50. Arrow KJ (1963) Uncertainty and the Welfare Economics of Medical-Care. American Economic Review 53: 941-973.

51. Morris, S., Devlin, N., and Parkin, D. (2007) Economic anaysis in health care. Chichester: John Wiley and Sons.

52. Sen, A. K. (1984) Resources, Values and Development. Oxford: Blackwell.

53. Sen, A. (1992) Inequality Reexamined. Oxford: Oxford University Press.

54. Sugden R (1993) Welfare, Resources, and Capabilities - A Review of Inequality Reexamined by Amartya Sen. Journal of Economic Literature 31: 1947-1962.

55. Sen, A. (1992) Inequality Reexamined. Oxford: Oxford University Press.

56. Nozick R (1973) Distributive Justice. Philosophy & Public Affairs 3: 45-126.

57. Rawls, J. (1971) A Theory of Justice. Cambridge: Havard University Press.

58. Sen, A. (1987) On Ethics and Economics. Oxford: Basil Blackwell Ltd.

59. Sen A (2002) Health: perception versus observation. BMJ 324: 860-861.

60. Sackett DL, Torrance GW (1978) The utility of different health states as perceived by the general public. Journal of Chronic Disease 31: 697-704.

61. Brickman P, Coates D, Janoffbulman R (1978) Lottery Winners and Accident Victims - Is Happiness Relative. Journal of Personality and Social Psychology 36: 917-927.

62. Oswald AJ, Powdthavee N (2008) Does happiness adapt? A longitudinal study of disability with implications for economists and judges. Journal of Public Economics 92: 1061-1077.

63. Ryan M, Shackley P (1995) Assessing the benefits of health care: how far should we go? Qual Health Care 4: 207-213.

64. Coast J (2009) Maximisation in extra-welfarism: A critique of the current position in health economics. Soc Sci Med 69: 786-792.

65. Olsen JA, Donaldson C (1998) Helicopters, hearts and hips: using willingness to pay to set priorities for public sector health care programmes. Soc Sci Med 46: 1-12.

66. Coast J, Smith R, Lorgelly P (2008) Should the capability approach be applied in health economics? Health Econ 17: 667-670.

67. Musgrave, R. A. (1959) The Economics of Public Finance. New York: McGraw Hill.

68. Tobin J (1970) On limiting the domain of inequality. Journal of Law and Economics 5: 263-278.

69. Walzer, M. (1983) Spheres of Justice. USA: Basic Books.

70. Sen A (1993) Capability and Well-Being. In: Nussbaum M, Sen A, editors. The Quality of Life. Oxford: Oxford University Press. pp. 30-53.

71. Sen A (2002) Why health equity? Health Economics 11: 659-666. 10.1002/hec.762.

72. Sugden R, Williams A (1978) The Objective in Cost-benefit Analysis. In: The Principles of Practical Cost-Benefit Analysis. Oxford: Oxford University Press. pp. 89-98.

73. Culyer AJ (1989) The Normative Economics of Health Finance and Provision. Oxford review of Economic Policy 5: 34-56.

74. Culyer AJ (1989) The Normative Economics of Health Finance and Provision. Oxford review of Economic Policy 5: 34-56.

75. Sen A (1980) Equality of what? In: Sen A, editors. Choice, Welfare and Measurement. Oxford: Blackwell.

76. Sen AK (1985) Well-being, Agency and Freedom: the Dewey Lectures 1984. Journal of Philosophy 82.

77. Birch S, Donaldson C (2003) Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? Soc Sci Med 56: 1121-1133.

78. Sen A (1980) Equality of what? In: Sen A, editors. Choice, Welfare and Measurement. Oxford: Blackwell.

79. Hurley J (1998) Welfarism, extra-welfarism and evaluative economic analysisin the health care sector. In: Barer ML, Getzen TE, Stoddard GL, editors. Health, Health Care and Health Economics: perspectives on distribution. Chichester: John Wiley & Sons Ltd.

80. Brouwer WBF, Koopmanschap MA (2000) On the economic foundations of CEA. Ladies and gentlemen, take your positions! Journal of Health Economics 19: 439-459.

81. Sen A (1993) Capability and Well-Being. In: Nussbaum M, Sen A, editors. The Quality of Life. Oxford: Oxford University Press. pp. 30-53.

82. Dolan P, Shaw R, Tsuchiya A, Williams A (2005) QALY maximisation and people's preferences: a methodological review of the literature. Health Economics 14: 197-208. 10.1002/hec.924.

83. Williams A (1997) Intergenerational equity: An exploration of the 'fair innings' argument. Health Economics 6: 117-132.

84. Theodorou M, Samara K, Pavlakis A, Middleton N, Polyzos N, Maniadakis N (2010) The public's and doctors' perceived role in participation in setting health care priorities in Greece. Hellenic Journal of Cardiology 51 (3) (pp 200-208), 2010 Date of Publication: May-June 2010 May-June.

85. Werner P (2009) Israeli lay persons' views on priority-setting criteria for Alzheimer's disease. Health Expectations 12: 187-196.

86. Alvarez B, Rodriguez-Miguez E (2011) Patients' self-interested preferences: Empirical evidence from a priority setting experiment. Social Science & Medicine 72: 1317-1324.

87. Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., and Stoddart, G. L. (2005) Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press.

88. Robinson R (1993) Cost-effectiveness analysis. BMJ 307: 793-795.

89. Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., and Stoddart, G. L. (2005) Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press.

90. Mehrez A, Gafni A (1989) Quality-adjusted life years, utility theory, and healthy-years equivalents. Med Decis Making 9: 142-149.

91. Torrance GW, Thomas W, Sackett DL (1972) A utility maximisation model for evaluation of healthcare programmes. Health Services Research 7: 118-133.

92. de Bekker-Grob EW, Ryan M, Gerard K (2012) Discrete choice experiments in health economics: a review of the literature. Health Econ 21: 145-172. 10.1002/hec.1697 [doi].

93. Ryan, M., Gerard, K., and Amaya-Amaya, M. (2008) Using Discrete Choice Experiments to Value Health and Health Care. Dordrecht, The Netherlands: Springer.

94. Ryan M, Gerard K (2003) Using discrete choice experiments to value health care programmes: current practice and future research reflections. Appl Health Econ Health Policy 2: 55-64.

95. Klarman HE, Francis JO, Rosenthal GD (1968) Cost Effectiveness Analysis Applied to Treatment of Chronic Renal Disease. Medical Care 6: 48-54.

96. Weinstein MC, Stason WB (1977) Foundations of Cost-Effectiveness Analysis for Health and Medical Practices. New England Journal of Medicine 296: 716-721.

97. Williams A (1985) Economics of coronary artery bypass grafting. Br Med J (Clin Res Ed) 291: 326-329.

98. Weinstein MC, Torrance G, McGuire A (2009) QALYs: The Basics. Value in Health 12: S5-S9.

99. Lorgelly PK, Lawson KD, Fenwick EA, Briggs AH (2010) Outcome measurement in economic evaluations of public health interventions: a role for the capability approach? Int J Environ Res Public Health 7: 2274-2289. 10.3390/ijerph7052274 [doi].

100. Coast J, Flynn T, Sutton E, Al-Janabi H, Vosper J, Lavender S, Louviere J, Peters T (2008) Investigating Choice Experiments for Preferences of Older People (ICEPOP): evaluative spaces in health economics. Journal of Health Services Research & Policy 13: 31-37.

101. Al-Janabi H, Flynn TN, Coast J (2011) QALYs and Carers. Pharmacoeconomics 29: 1015-1023.

102. Goodrich K, Kaambwa B, Al-Janabi H (2012) The Inclusion of Informal Care in Applied Economic Evaluation: A Review. Value in Health 15: 975-981.

103. Fanshel S, Bush JW (1970) Health-Status Index and Its Application to Health-Services Outcomes. Operations Research 18: 1021-&.

104. Baker R, Bateman I, Donaldson C, Jones-Lee M, Lancsar E, Loomes G, Mason H, Odejar M, Prades JLP, Robinson A, Ryan M, Shackley P, Smith R, Sugden R, Wildman J (2010) Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project. Health Technology Assessment 14: 1-+.

105. Green C, Gerard K (2009) Exploring the Social Value of Health-Care Interventions: A Stated Preference Discrete Choice Experiment. Health Economics 18: 951-976.

106. Schomerus G, Matschinger H, Angermeyer MC (2006) Alcoholism: Illness beliefs and resource allocation preferences of the public. Drug and Alcohol Dependence 82: 204-210.

107. Nussbaum MC (2003) Capabilities as fundamental entitlements: Sen and social justice. Feminist Economics 9: 33-59. DOI 10.1080/1354570022000077926.

108. Sen, A. K. (1982) Choice, welfare and measurement. Cambridge, MA: Harvard University Press.

109. Grewal I, Lewis J, Flynn T, Brown J, Bond J, Coast J (2006) Developing attributes for a generic quality of life measure for older people: Preferences or capabilities? Social Science & Medicine 62: 1891-1901. DOI 10.1016/j.socscimed.2005.08.023.

110. Mitra S (2006) The Capability Approach and Disability. Journal of Disability Policy Studies 16: 236-247.

111. Robeyns I (2006) The capability approach in practice. Journal of Political Philosophy 14: 351-376.

112. Burchardt T (2004) Capabilities and disability: the capabilities framework and the social model of disability. Disability & Society 19: 735-751.

113. Lorgelly P, Lorimer K, Fenwick EA, Briggs AH (2008) The Capability Approach: developing an instrument for evaluating public health interventions.

114. Al-Janabi H, Keeley T, Mitchell P, Coast J (2013) Completion of self-report health and capability measures: a think aloud study. Soc Sci Med (accepted) .

115. Al-Janabi H, Flynn TN, Coast J (2012) Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. Quality of Life Research 21: 167-176.

116. Al-Janabi H, Peters TJ, Brazier J, Bryan S, Flynn TN, Clemens S, Moody A, Coast J (2012) An investigation of the construct validity of the ICECAP-A capability measure. Qual Life Res . 10.1007/s11136-012-0293-5 [doi].

117. Coast J, Peters TJ, Natarajan L, Sproston K, Flynn T (2008) An assessment of the construct validity of the descriptive system for the ICECAP capability measure for older people. Quality of Life Research 17: 967-976. DOI 10.1007/s11136-008-9372-z.

118. Coast J, Flynn TN, Natarajan L, Sproston K, Lewis J, Louviere JJ, Peters TJ (2008) Valuing the ICECAP capability index for older people. Social Science & Medicine 67: 874-882.

119. Friedman, L. M., Furberg, C. D., and DeMets, D. L. (2010) Fundamentals of Clinical Trials. New York: Springer.

120. Petrou S, Gray A (2011) Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. British Medical Journal 342.

121. Clauser SB, Ganz PA, Lipscomb J, Reeve BB (2007) Patient-reported outcomes assessment in cancer trials: Evaluating and enhancing the payoff to decision making. Journal of Clinical Oncology 25: 5049-5050.

122. Lipscomb J, Reeve BB, Clauser SB, Abrams JS, Bruner DW, Burke LB, Denicoff AM, Ganz PA, Gondek K, Minasian LM, O'Mara AM, Revicki DA, Rock EP, Rowland JH, Sgambati M, Trimble EL (2007) Patient-reported outcomes assessment in cancer trials: Taking stock, moving forward. Journal of Clinical Oncology 25: 5133-5140.

123. Lilienfeld AM (1982) The Fielding H. Garrison Lecture: Ceteris paribus: the evolution of the clinical trial. Bull Hist Med 56: 1-18.

124. Altman DG, Bland JM (1999) Statistics notes - Treatment allocation in controlled trials: why randomise? British Medical Journal 318: 1209.

125. Lewis JA (1999) Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline. Statistics in Medicine 18: 1903-1904.

126. Friedman LM, Furberg CD, DeMets DL (2010) Basic Study Design. In: Fundamentals of Clinical Trials. New York: Springer. pp. 67-96.

127. Friedman LM, Furberg CD, DeMets DL (2010) The randomization process. In: Fundamentals of Clinical Trials. New York: Springer. pp. 97-117.

128. Hill AB (1951) The Clinical Trial. British Medical Bulletin 7: 278-282.

129. Pocock SJ (1979) Allocation of Patients to Treatment in Clinical-Trials. Biometrics 35: 183-197.

130. Zelen M (1974) Randomization and Stratification of Patients to Clinical Trials. Journal of Chronic Diseases 27: 365-375.

131. Hrobjartsson A, Boutron I (2011) Blinding in Randomized Clinical Trials: Imposed Impartiality. Clinical Pharmacology & Therapeutics 90: 732-736.

132. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D (2002) Randomised trials in surgery: problems and possible solutions. British Medical Journal 324: 1448-1451.

133. Sedgwick P (2011) Statistical Question Superiority Trials. British Medical Journal 342.

134. Sedgwick P (2011) Statistical Question: Non-inferiority trials. British Medical Journal 342.

135. Logan AG, Milne BJ, Achber C, Campbell WP, Haynes RB (1981) Cost-Effectiveness of A Worksite Hypertension Treatment Program. Hypertension 3: 211-218.

136. Hull R, Hirsh J, Sackett DL, Stoddart G (1981) Cost-Effectiveness of Clinical-Diagnosis, Venography, and Non-Invasive Testing in Patients with Symptomatic Deep-Vein Thrombosis. New England Journal of Medicine 304: 1561-1567.

137. Sculpher MJ, Buxton MJ (1993) The Episode-Free Day As A Composite Measure of Effectiveness - An Illustrative Economic-Evaluation of Formoterol Versus Salbutamol in Asthma Therapy. Pharmacoeconomics 4: 345-352.

138. Fayers, P. M. and Machin, D. (2000) Quality of Life: assessment, analysis and interpretation. Chicester: John Wiley and Sons Ltd.

139. Karnofsky DA, Abelmann WH, Craver LF, Burchenal JH (1948) The Use of the Nitrogen Mustards in the Palliative Treatment of Carcinoma - with Particular Reference to Bronchogenic Carcinoma. Cancer 1: 634-656.

140. Bergner M, Bobbitt RA, Carter WB, Gilson BS (1981) The Sickness Impact Profile - Development and Final Revision of A Health-Status Measure. Medical Care 19: 787-805.

141. Hunt SM, Mckenna SP, Mcewen J, Williams J, Papp E (1981) The Nottingham Health Profile - Subjective Health-Status and Medical Consultations. Social Science & Medicine Part A-Medical Sociology 15: 221-229.

142. Brooks RG, Jendteg S, Lindgren B, Persson U, Bjork S (1991) EuroQol: health-related quality of life measurement. Results of the Swedish questionnaire exercise. Health Policy 18: 37-48.

143. Brooks R (1996) EuroQol: The current state of play. Health Policy 37: 53-72.

144. Ware JE, Jr., Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30: 473-483.

145. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, Dehaes JCJM, Kaasa S, Klee M, Osoba D, Razavi D, Rofe PB, Schraub S, Sneeuw K, Sullivan M, Takeda F (1993) The European-Organization-For-Research-And-Treatment-Of-Cancer Qlq-C30 - A Quality-Of-Life Instrument for Use in International Clinical-Trials in Oncology. Journal of the National Cancer Institute 85: 365-376.

146. EuroQol (2012) http://www.euroqol.org/eq-5d-products/eq-5d-5l.html.

147. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ (2008) Comparing the standard EQ-5D three-level system with a five-level version. Value in Health 11: 275-284.

148. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, Swinburn P, Busschbach J (2012) Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res . 10.1007/s11136-012-0322-4 [doi].

149. European Organisation for Research and Treatment of Cancer (2013) EORTC QLQ-C30 modules.

150. Zigmond AS, Snaith RP (1983) The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandinavica 67: 361-370.

151. Varni JW, Seid M, Rode CA (1999) The PedsQL (TM): Measurement model for the pediatric quality of life inventory. Medical Care 37: 126-139.

152. Parmelee PA, Katz IR (1990) Geriatric Depression Scale. Journal of the American Geriatrics Society 38: 1379.

153. Brazier J, Roberts J, Deverill M (2002) The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 21: 271-292.

154. Brazier J, Roberts J, Tsuchiya A, Busschbach J (2004) A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics 13: 873-884. DOI 10.1002/hec.866.

155. Furlong WJ, Feeny DH, Torrance GW, Barr RD (2001) The Health Utilities Index (HUI (R)) system for assessing health-related quality of life in clinical studies. Annals of Medicine 33: 375-384.

156. Horsman JR, Fluchel M, Furlong W, Castillo L, Barr RD (2003) Agreement of Health Utilities Index scores among survivors of cancer in childhood their parents and their doctors in Uruguay. Value in Health 6: 233.

157. Dolan P, Gudex C, Kind P, Williams A (1996) The time trade-off method: Results from a general population study. Health Economics 5: 141-154.

158. Anand P, Hunter G, Smith R (2005) Capabilities and well-being: Evidence based on the Sen-Nussbaum approach to welfare. Social Indicators Research 74: 9-55.

159. Anand P, Hunter G, Carter I, Dowding K, Guala F, Van Hees M (2009) The Development of Capability Indicators. Journal of Human Development and Capabilities 10: 125-152. DOI 10.1080/14649880802675366.

160. Simon J, Anand P, Gray A, Rugkasa J, Yeeles K, Burns T (2013) Operationalising the capability approach for outcome measurement in mental health research. Soc Sci Med 98: 187-196.

161. Netten A, Ryan M, Smith P, Skatum D, Healey A, Knapp M (2002) The development of a measure of social care outcome for older people.

162. Netten A, Hirst M, Glendinning C (2005) Developing a measure of PSS outputs: interim report.

163. Forder JE, Caiels J (2011) Measuring the outcomes of long-term care. Social Science & Medicine 73: 1766-1774.

164. The University of Birmingham (2013) ICECAP Capability Measures.

165. Flynn TN, Louviere JJ, Peters TJ, Coast J (2007) Best--worst scaling: What it can do for health care research and how to do it. J Health Econ 26: 171-189. S0167-6296(06)00049-X [pii];10.1016/j.jhealeco.2006.04.002 [doi].

166. Flynn T, Huynh E, Peters T, Al-Janabi H, Clemens S, Moody A, Coast J (2013) Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. Health Economics .

167. National Institute for Health and Care Excellence (2013) NICE and social care.

168. National Institute for Health and Care Excellence (2013) NICE announces Collaborating Centre for Social Care.

169. National Institute for Health and Care Excellence (2013) The social care guidance manual.

170. Social Care Institute for Excellence (2013) SCIE's approach to economic evaluation in social care. Adults' services SCIE report 52.

171. CRD UoY (2008) Systematic Reviews: CRD's guidance for undertaking reviews in health care.

172. Makai P, Brouwer WB, Koopmanschap MA, Nieboer AP (2012) Capabilities and quality of life in Dutch psycho-geriatric nursing homes: an exploratory study using a proxy version of the ICECAP-O. Qual Life Res 21: 801-812. 10.1007/s11136-011-9997-1 [doi].

173. Davis JC, Liu-Ambrose T, Richardson CG, Bryan S (2012) A comparison of the ICECAP-O with EQ-5D in a falls prevention clinical setting: are they complements or substitutes? Qual Life Res . 10.1007/s11136-012-0225-4 [doi].

174. Davis JC, 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 Geriatr 12: 65. 1471-2318-12-65 [pii];10.1186/1471-2318-12-65 [doi].

175. Couzner L, Ratcliffe J, Crotty M (2012) The relationship between quality of life, health and care transition: an empirical comparison in an older post-acute population. Health Qual Life Outcomes 10: 69. 1477-7525-10-69 [pii];10.1186/1477-7525-10-69 [doi].

176. Ratcliffe J, Laver K, Couzner L, Quinn S, Crotty M (2011) An assessment of the construct validity of the ICECAP-O indexof capability in Australian national transition care and clinical rehabilitation programmes.

177. Flynn TN, Chan P, Coast J, Peters TJ (2011) Assessing quality of life among British older people using the ICEPOP CAPability (ICECAP-O) measure. Appl Health Econ Health Policy 9: 317-329. 3 [pii];10.2165/11594150-000000000-00000 [doi].

178. Couzner L, Ratcliffe J, Lester L, Flynn T, Crotty M (2013) Measuring and valuing quality of life for public health research: application of the ICECAP-O capability index in the Australian general population. International Journal of Public Health 58: 367-376.

179. Ratcliffe J, Lester LH, Couzner L, Crotty M (2013) An assessment of the relationship between informal caring and quality of life in older community-dwelling adults - more positives than negatives? Health & Social Care in the Community 21: 35-46.

180. Makai P, Koopmanschap MA, Brouwer WBF, Nieboer AAP (2013) A validation of the ICECAP-O in a population of post-hospitalized older people in the Netherlands. Health and Quality of Life Outcomes 11.

181. Al-Janabi H, Keeley T, Mitchell P, Coast J (2013) Can capabilities be self-reported? A think aloud study. Social Science & Medicine 87: 116-122.

182. Horwood J, Sutton E, Coast J (2013) Evaluating the Face Validity of the ICECAP-O Capabilities Measure: A "Think Aloud" Study with Hip and Knee Arthroplasty Patients. Applied Research Quality of Life .

183. Williamson P, Altman D, Blazeby J, Clarke M, Gargon E (2012) Driving up the quality and relevance of research through the use of agreed core outcomes. Journal of Health Services Research & Policy 17: 1-2.

184. Streiner, D. L. and Norman, G. R. (2008) Health Measurement Scales: a practical guide to their development and use. Oxford: Oxford University Press.

185. American Educational Research Association American Psychological Association National Council on Measurement in Education (1999) Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

186. Streiner, D. L. and Norman, G. R. (2008) Health Measurement Scales: a practical guide to their development and use. Oxford: Oxford University Press.

187. Stevens SS (1951) Mathematics, Measurement and Psychophysics.

188. Carmines, E. G. and Zeller, R. A. (1979) Reliability and Validity Assessment. London: Sage Publications.

189. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD, Sloan JA (2007) What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value in Health 10: S94-S105. DOI 10.1111/j.1524-4733.2007.00272.x.

190. Stanley JC (1971) Reliability. In: Thorndike RL, editors. Educational Measurement. Washington DC. pp. 356-442.

191. Carmines, E. G. and Zeller, R. A. (1979) Reliability and Validity Assessment. London: Sage Publications.

192. Gronlund, N. E. and Lin, R. L. (1990) Measurement and Evaluation in teaching. New York: Macmillan.

193. Helms JE, Henze KT, Sass TL, Mifsud VA (2006) Treating Cronbach's alpha reliability coefficients as data in counseling research. Counseling Psychologist 34: 630-660.

194. Brazier, J., Ratcliffe, J., Salomon, J. A., and Tsuchiya, A. (2007) Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press.

195. Nunnally, J. C. and Bernstein, I. H. (1994) Psychometric theory. New York: McGraw-Hill.

196. Waltz, C. F., Strickland, O. L., and Lenz, E. R. (2005) Measurement in nursing and health research. New York: Springer.

197. Cronbach LJ (1951) Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16: 297-334.

198. Colliver JA, Conlee MJ, Verhulst SJ (2012) From test validity to construct validity ... and back? Medical Education 46: 366-371.

199. Messick S (1989) Validity. In: Linn RL, editors. Educational Measurement. New York: Macmillan. pp. 13-103.

200. Kane MT (2001) Current concerns in validity theory. Journal of Educational Measurement 38: 319-342.

201. Guilford JP (1946) New Standards for Test Evaluation. Educational and Psychological Measurement 6: 427-439.

202. Landy FJ (1986) Stamp Collecting Versus Science - Validation As Hypothesis-Testing. American Psychologist 41: 1183-1192.

203. Cronbach LJ, Meehl PE (1955) Construct Validity in Psychological Tests. Psychological Bulletin 52: 281-302.

204. Landy FJ (1986) Stamp Collecting Versus Science - Validation As Hypothesis-Testing. American Psychologist 41: 1183-1192.

205. Kane MT (1992) An Argument-Based Approach to Validity. Psychological Bulletin 112: 527-535.

206. Messick S (1989) Validity. In: Linn RL, editors. Educational Measurement. New York: Macmillan. pp. 13-103.

207. Messick S (1980) Test Validity and the Ethics of Assessment. American Psychologist 35: 1012-1027.

208. Loevinger J (1957) Objective Tests As Instruments of Psychological Theory. Psychological Reports 3: 635-694.

209. Borsboom D, Mellenbergh GJ, van Heerden J (2004) The concept of validity. Psychological Review 111: 1061-1071.

210. Borsboom D (2006) The attack of the psychometricians. Psychometrika 71: 425-440.

211. Borsboom D, Cramer AOJ, Kievit RA, Scholten AZ, Franic S (2009) The End of construct validity. In: Lissitz RW, editors. The ceoncept of validity. Charlotte: Information Age. pp. 135-170.

212. Kane MT (2009) Validating the interpretations and uses of test scores. In: Lissitz RW, editors. The concept of validity. Charlotte, NC: Information Age Publishing. pp. 39-64.

213. Lawshe CH (1985) Inferences from Personnel Tests and Their Validity. Journal of Applied Psychology 70: 237-238.

214. Downing SM (2003) Validity: on the meaningful interpretation of assessment data. Medical Education 37: 830-837.

215. Hays RD, Hadorn D (1992) Responsiveness to Change: An Aspect of Validity, Not a Separate Dimension. Quality of Life Research 1: 73-75.

216. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD, Sloan JA (2007) What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value in Health 10: S94-S105. DOI 10.1111/j.1524-4733.2007.00272.x.

217. Sireci SG (2009) Packing and Unpacking Sources of Validity Evidence. In: Lissitz RW, editors. The Concept of validity. Charlotte, NC: Information Age Publishing. pp. 19-37.

218. Cronbach LJ (1971) Test Validation. In: Thorndike RL, editors. Educational Measurement. Washington, DC: American Council on Education. pp. 443-507.

219. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, De Vet HCW (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology 63: 737-745.

220. Fransen M, Edmonds J (1999) Reliability and validity of the EnroQol in patients with osteoarthritis of the knee. Rheumatology 38: 807-813.

221. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A (1997) Measuring health-related quality of life in rheumatoid arthritis: Validity, responsiveness and reliability of EuroQol (EQ-5D). British Journal of Rheumatology 36: 551-559.

222. Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H (1994) Validity of Euroqol - A Generic Health-Status Instrument in Patients with Rheumatoid-Arthritis. British Journal of Rheumatology 33: 655-662.

223. Schrag A, Selai C, Jahanshahi M, Quinn NP (2000) The EQ-5D - a generic quality of life measure - is a useful instrument to measure quality of life in patients with Parkinson's disease. Journal of Neurology Neurosurgery and Psychiatry 69: 67-73.

224. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Hetland ML (2008) Health-related quality of life: Validity, reliability, and responsiveness of SF-36, EQ-15D, EQ-5D, RAQoL, and HAQ in patients with rheumatoid arthritis. Journal of Rheumatology 35: 1528-1537.

225. Lawshe CH (1985) Inferences from Personnel Tests and Their Validity. Journal of Applied Psychology 70: 237-238.

226. Kane MT (2001) Current concerns in validity theory. Journal of Educational Measurement 38: 319-342.

227. Cronbach LJ, Meehl PE (1955) Construct Validity in Psychological Tests. Psychological Bulletin 52: 281-302.

228. Coast J, Peters TJ, Natarajan L, Sproston K, Flynn T (2008) An assessment of the construct validity of the descriptive system for the ICECAP capability measure for older people. Quality of Life Research 17: 967-976. DOI 10.1007/s11136-008-9372-z.

229. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, Schwartz C, Revicki DA, Moinpour CM, McLeod LD, Lyons JC, Lenderking WR, Hinds PS, Hays RD, Greenhalgh J, Gershon R, Feeny D, Fayers PM, Cella D, Brundage M, Ahmed S, Aaronson NK, Butt Z (2013) ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res . 10.1007/s11136-012-0344-y [doi].

230. Westen D, Rosenthal R (2003) Quantifying construct validity: Two simple measures. Journal of Personality and Social Psychology 84: 608-618.

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

232. Brod M, Tesler LE, Christensen TL (2009) Qualitative research and content validity: developing best practices based on science and experience. Quality of Life Research 18: 1263-1278. DOI 10.1007/s11136-009-9540-9.

233. Magasi S, Ryan G, Revicki D, Lenderking W, Hays RD, Brod M, Snyder C, Boers M, Cella D (2012) Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Quality of Life Research 21: 739-746.

234. Smith GT (2005) On construct validity: Issues of method and measurement. Psychological Assessment 17: 396-408.

235. Guion RM (1980) On trinitarian doctrines of validity. Professional Psychology 11: 385-398.

236. Kane MT (1994) Validating Interpretive Arguments for Licensure and Certification Examinations. Evaluation & the Health Professions 17: 133-159.

237. Mosier CL (1947) A Critical Examination of the Concepts of Face Validity. Educational and Psychological Measurement 7: 191-205.

238. Denzin, N. K. and Lincoln, Y. S. (2000) Handbook of Qualitative Research. Thousand Oaks, CA: Sage.

239. Glaser, B. G. and Strauss, A. L. (1967) The Discovery of Grounded Theory: Startegies for Qualitative Research. Chicago: Aldine de Gruyter.

240. Ritchie J, Spencer L, O'Connor W (2009) Carrying out Qualitative Analysis. In: Ritchie J, Lewis J, editors. Qualitative Research Practice. London: Sage. pp. 219-262.

241. Spencer L, Ritchie J, O'Connor W (2009) Analysis: Practices, Principles and Processes. In: Ritchie J, Lewis J, editors. Qualitative Research Practice. London: SAGE Publications. pp. 199-218.

242. Patton, M. Q. (2002) Qualitative Research and Evaluation methods. Thousand Oaks, CA: Sage.

243. Mason, J. (2002) Qualitative Researching. London.

244. Ritchie J (2009) The Applications of Qualitative Methods to Social Research. In: Ritchie J, Lewis J, editors. Qualitative Research Practice. London: Sage. pp. 25-46.

245. Badia X, Monserrat S, Roset M, Herdman M (1999) Feasibility, validity and test-retest reliability of scaling methods for health states: The visual analogue scale and the time trade-off. Quality of Life Research 8: 303-310.

246. Laing MH (2000) Longitudinal Construct Validity: Establishment of Clinical Meaning in Patient Evaluative Instruments. Medical Care 38.

247. Laing MH (2000) Longitudinal Construct Validity: Establishment of Clinical Meaning in Patient Evaluative Instruments. Medical Care 38.

248. Guyatt G, Walter S, Norman G (1987) Measuring Change Over Time - Assessing the Usefulness of Evaluative Instruments. Journal of Chronic Diseases 40: 171-178.

249. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of Health-Status - Ascertaining the Minimal Clinically Important Difference. Controlled Clinical Trials 10: 407-415.

250. King MT (2011) A point of minimal important difference (MID): a critique of terminology and methods. Expert Review of Pharmacoeconomics & Outcomes Research 11: 171-184.

251. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR (2002) Methods for explaining the clinical significance of health status measures. Clinical Therapeutics 24: 5.

252. Schunemann HJ, Guyatt GH (2005) Commentary - Goodbye M(C)ID! Hello MID, where do you come from? Health Services Research 40: 593-597.

253. Guyatt G, Walter S, Norman G (1987) Measuring Change Over Time - Assessing the Usefulness of Evaluative Instruments. Journal of Chronic Diseases 40: 171-178.

254. Hays RD, Hadorn D (1992) Responsiveness to Change: An Aspect of Validity, Not a Separate Dimension. Quality of Life Research 1: 73-75.

255. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LA, De Vet HCW (2007) Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 60: 34-42.

256. McHorney CA, Tarlov AR (1995) Individual-Patient Monitoring in Clinical-Practice - Are Available Health-Status Surveys Adequate. Quality of Life Research 4: 293-307.

257. Brazier J, Jones N, Kind P (1993) Testing the Validity of the Euroqol and Comparing It with the Sf-36 Health Survey Questionnaire. Quality of Life Research 2: 169-180.

258. Brazier J, Roberts J, Tsuchiya A, Busschbach J (2004) A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics 13: 873-884. DOI 10.1002/hec.866.

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

260. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S (2012) Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res . 10.1007/s11136-012-0175-x [doi].

261. Revicki D, Hays RD, Cella D, Sloan J (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 61: 102-109. S0895-4356(07)00119-9 [pii];10.1016/j.jclinepi.2007.03.012 [doi].

262. Walters SJ, Brazier JE (2005) Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 14: 1523-1532.

263. Yost KJ, Eton DT (2005) Combining distribution- and anchor-based approaches to determine minimally important differences: The FACIT experience. Evaluation & the Health Professions 28: 172-191.

264. Brozek JL, Guyatt GH, Schunemann HJ (2006) How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health and Quality of Life Outcomes 4.

265. Sireci SG (1998) The construct of content validity. Social Indicators Research 45: 83-117.

266. Lasch KE, Marquis P, Vigneux M, Abetz L, Arnould B, Bayliss M, Crawford B, Rosa K (2010) PRO development: rigorous qualitative research as the crucial foundation. Quality of Life Research 19: 1087-1096.

267. Strauss, A. L. and Corbin, J. (1990) Basics of qualitative research. California, US: Sage.

268. Coast J, McDonald R, Baker R (2004) Issues arising from the use of qualitative methods in health economics. J Health Serv Res Policy 9: 171-176. 10.1258/1355819041403286 [doi].

269. Green, J. and Thorogood, N. (2009) Qualitative Methods for Health Research. London, UK: Sage.

270. Snape D, Spencer L (2009) The Foundations of Qualitative Research. In: Ritchie J, Lewis J, editors. Qualitative Research Practice: A guide for social science students and researchers. London: Sage. pp. 1-23.

271. Ritchie J, Lewis J, Elam G (2009) Designing ans Selecting Samples. In: Ritchie J, Lewis J, editors. Qualitative Research Practice. London: SAGE Publications. pp. 77-108.

272. Miles, M. B. and Huberman, A. M. (1994) Qualitative Data Analysis: AN Expanded Sourcebook. London: Sage.

273. Hammersley, M. and Atkinson, P. (1995) Ethnography: principles and Practice. Routledge: London.

274. Silverman D (2000) Analysing conversation. In: Seale C, editors. Researching Society and Culture. London: Sage.

275. Tonkiss F (2000) Analysing discourse. In: Seale C, editors. Researching Society and Culture. London: Sage.

276. Robinson WS (1951) The Logical Structure of Analytic Induction. American Sociological Review 16: 812-818.

277. Glaser BG (1965) The Constant Comparative Method of Qualitative-Analysis. Social Problems 12: 436-445.

278. Legard R, Keegan J, Ward K (2009) In-depth Interviews. In: Ritchie J, Lewis J, editors. Qualitative Research Practice. London: SAGE Publishing. pp. 138-169.

279. Stroke Prevention Progamme PCtUoB (2010) Past BP: A randomised controlled trial of different blood pressure targets for people with a history of stroke and transient ischaemic attack (TIA) in primary care. 4.

280. Keele University (2010) Improving the effectiveness of exercise for knee pain in older adults in primary care: Benefits of Effective Exercise for knee Pain (BEEP). 2.

281. Bellamy N (2002) WOMAC Osteoarthritis Index User Guide. Version V.

282. Essink-Bot ML, Bonsel GJ, van der Maas PJ (1990) Valuation of health states by the general public: feasibility of a standardized measurement procedure. Soc Sci Med 31: 1201-1206.

283. 1990) EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 16: 199-208.

284. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X (2011) Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research 20: 1727-1736.

285. Harrison MJ, Davies LM, Bansback NJ, Ingram M, Anis AH, Symmons DP (2008) The validity and responsiveness of generic utility measures in rheumatoid arthritis: a review. J Rheumatol 35: 592-602. 08/13/028 [pii].

286. Papaioannou D, Brazier J, Parry G (2011) How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value Health 14: 907-920. S1098-3015(11)01415-X [pii];10.1016/j.jval.2011.04.006 [doi].

287. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A (1997) Measuring health-related quality of life in rheumatoid arthritis: Validity, responsiveness and reliability of EuroQol (EQ-5D). British Journal of Rheumatology 36: 551-559.

288. Maruish ME (2011) User's manual for the SF-36v2 health survey. Third Edition.

289. Hawthorne G, Osborne RH, Taylor A, Sansoni J (2007) The SF36 Version 2: critical analyses of population weights, scoring algorithms and population norms. Quality of Life Research 16: 661-673.

290. Brazier JE, Harper R, Jones NMB, Ocathain A, Thomas KJ, Usherwood T, Westlake L (1992) Validating the Sf-36 Health Survey Questionnaire - New Outcome Measure for Primary Care. British Medical Journal 305: 160-164.

291. Bellamy N, Campbell J, Hill J, Band P (2002) A comparative study of telephone versus onsite completion of the WOMAC 3.0 Osteoarthritis Index. Journal of Rheumatology 29: 783-786.

292. Theiler R, Spielberger J, Bischoff HA, Bellamy N, Huber J, Kroesen S (2002) Clinical evaluation of the WOMAC 3.0 OA Index in numeric rating scale format using a computerized touch screen version. Osteoarthritis Cartilage 10: 479-481. 10.1053/joca.2002.0807 [doi];S1063458402908071 [pii].

293. Soderman P, Malchau H (2000) Validity and reliability of Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). Acta Orthop Scand 71: 39-46. 10.1080/00016470052943874 [doi].

294. Broadbent E, Petrie KJ, Main J, Weinman J (2006) The brief illness perception questionnaire. J Psychosom Res 60: 631-637. S0022-3999(05)00491-5 [pii];10.1016/j.jpsychores.2005.10.020 [doi].

295. Spitzer RL, Kroenke K, Williams JBW, Lowe B (2006) A brief measure for assessing generalized anxiety disorder - The GAD-7. Archives of Internal Medicine 166: 1092-1097.

296. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH (2009) The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders 114: 163-173.

297. Lowe B, Decker O, Muller S, Brahler E, Schellberg D, Herzog W, Herzberg PY (2008) Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. Medical Care 46: 266-274.

298. Farrell B, Godwin J, Richards S, Warlow C (1991) The United-Kingdom Transient Ischemic Attack (Uk-Tia) Aspirin Trial - Final Results. Journal of Neurology Neurosurgery and Psychiatry 54: 1044-1054.

299. Wilson JTL, Hareendran A, Hendry A, Potter J, Bone I, Muir KW (2005) Reliability of the modified rankin scale across multiple raters - Benefits of a structured interview. Stroke 36: 777-781.

300. Sedgwick P (2012) Pearson's correlation coefficient. British Medical Journal 344.

301. Sedgwick P (2012) Correlation. British Medical Journal 345.

302. Goodman LA, Kruskal WH (1954) Measures of Association for Cross Classifications. Journal of the American Statistical Association 49: 732-764.

303. Bland, M. (2000) An introduction to medical statistics. Oxford: Oxford University Press.

304. Kaiser HF (1960) The Application of Electronic-Computers to Factor-Analysis. Educational and Psychological Measurement 20: 141-151.

305. Hendrickson AE, White PO (1964) Promax - A Quick Method for Rotation to Oblique Simple Structure. British Journal of Statistical Psychology 17: 65-70.

306. Brown JD (2009) Choosing the right type of rotation in PCA and EFA. JALT Testing & Evaluation SIG Newsletter 13: 20-25.

307. Husted JA, Cook RJ, Farewell VT, Gladman DD (2000) Methods for assessing responsiveness: a critical review and recommendations. Journal of Clinical Epidemiology 53: 459-468.

308. Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and Responsiveness of Health-Status Measures - Statistics and Strategies for Evaluation. Controlled Clinical Trials 12: S142-S158.

309. Cohen, J. (1988) Statistical power analysis for the behavioural sciences. New york: Psychology Press.

310. Sivan M (2009) Interpreting Effect Size to Estimate Responsiveness of Outcome Measures. Stroke 40: E709.

311. Kind P, Hardman G, Macran S (1993) UK Population Norms for EQ-5D.

312. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Lowe B (2007) Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. Annals of Internal Medicine 146: 317-325.

313. Alzheimer's Society (2012) The Mini Mental State Examination (MMSE). 1-5.

314. Tracy SJ (2010) Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research. Qualitative Inquiry 16: 837-851.

315. Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, Grimshaw JM (2010) What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychology and Health 25: 1229-1245.

316. Guest G, Bunce A, Johnson L (2006) How many interviews are enough? An experiment with data saturation and variability. Field Methods 18: 59-82.

317. Kim TH, Jo MW, Lee SI, Kim SH, Chung SM (2012) Psychometric properties of the EQ-5D-5L in the general population of South Korea. Qual Life Res . 10.1007/s11136-012-0331-3 [doi].

318. Birch S, Donaldson C (2003) Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? Soc Sci Med 56: 1121-1133.

319. Robinson R (1993) Cost-utility analysis. BMJ 307: 859-862.

320. Sen A (2002) Health: perception versus observation. BMJ 324: 860-861.

321. Cooper R, Kuh D, Cooper C, Gale CR, Lawlor DA, Matthews F, Hardy R (2011) Objective measures of physical capability and subsequent health: a systematic review. Age and Ageing 40: 14-23.

322. Pickard AS, Neary MP, Cella D (2007) Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes 5: 70. 1477-7525-5-70 [pii];10.1186/1477-7525-5-70 [doi].