The measurement of locomotor disability in epidemiological studies

Sara Nicole Muller

A thesis submitted for the degree of

Doctor of Philosophy

September 2010

Arthritis Research UK Primary Care Centre

Keele University

Declaration

This PhD project was nested within the North Staffordshire Osteoarthritis Project (NorStOP), which was funded by a Medical Research Council (MRC) programme grant (grant code: G9900220 and G0501798) obtained at the Arthritis Research UK Primary Care Centre by Professor Peter Croft and colleagues. The idea for the PhD project formed part of the submission to the MRC for this grant.

Throughout the course of the PhD project, with guidance from my supervisors Dr Elaine Thomas and Dr George Peat, I developed the ideas around and managed the direction of the thesis.

The NorStOP baseline data were collected prior to my involvement in the study. In my capacity as Research Assistant in Biostatistics at the Arthritis Research UK Primary Care Centre, I had some involvement in the data cleaning of the NorStOP1 three and six-year follow-up datasets. Dr Elaine Thomas and Ms Tracy Whitehurst provided me with background information on the NorStOP study, the procedures used to collect data and measures taken to enhance response rates. In this thesis, I have approached the use of the NorStOP data as a secondary analysis of previously collected data. I had no control over the content of the surveys.

I obtained data from the Welsh Health Survey through the UK Data Archive. Dr Clare Jinks provided data from the Keele Knee Pain Cohort Study and Professor Danielle van der Windt provided data from a Dutch cohort.

Dr Elaine Thomas and Dr George Peat advised me on the planning of all analyses and on the writing and presentation of chapters. I conducted all analyses and wrote the chapters myself. I received guidance on the search strategy for the literature review from Dr Kate Dunn and Ms Joanne Jordan. Dr Richard Williams (University of Notre Dame) provided advice by email on the ordinal modelling aspects of the thesis, in particular the use of the -gologit2- program. Dr David Andrich (University of Western Australia), Dr Svetlana Beltyukova (University of Toledo), Dr Barry Sheridan (University of Western Australia) and Prof Alan Tennant (University of Leeds) provided advice by email and in person on various aspects of the Rasch model, it applications and limitations.

Abstract

This thesis is concerned with the measurement of locomotor disability (LMD) in epidemiological studies. The central hypothesis was that LMD is a continuous phenomenon and research into this important health indicator, with specific reference to its relationship to pain in community-dwelling adults aged 50+ years, could be improved by interval-level measurement, rather than binary definitions.

A systematic search and narrative review of the literature revealed a range of concepts and content of previous self-complete LMD instruments, and an absence of interval-level measures. A brief, self-complete scale of physical functioning, the PF-10, commonly used in epidemiological studies, and suggested as a measure of LMD, was taken as the starting point for empirical work in this thesis.

A subset of five items mapped onto the LMD construct and possessed acceptable psychometric properties.

Analysis of cross-sectional data from 18,497 adults using ordinal regression models and individual item responses illustrated one, albeit relatively inefficient, approach to moving beyond binary outcomes for investigating the association between pain and LMD.

An interval-level measure of LMD was derived using the Rasch model and combining the five items into two super-items (walking, stair-climbing). The scoring mechanism was externally verified in local, national and international datasets, and the psychometric properties confirmed.

Data from 680 initially pain-free adults were used to demonstrate the potential of the new measure for longitudinal analysis. This suggested a right-shift (worsening) in the distribution of LMD at three and six years. Pain onset resulted in a more rapid increase in LMD, and recovery from pain led to only a partial return to pre-pain levels.

Locomotor disability exists on a continuum and its measurement should reflect this. An interval-level measure was derived from a set of commonly used items. This measure offers several advantages (brevity, application to retrospectively gathered data) but also has limitations (ceiling/floor effects).

Acknowledgements

I would like to express my thanks to several people who have made the writing of this thesis possible.

First and foremost my supervisors Dr Elaine Thomas and Dr George Peat for their invaluable advice and comments on the drafts of this thesis. Also, the Arthritis Research UK Primary Care Centre for funding my time as a student at Keele and to all the staff and students there for their support.

In addition, the experts on various email interest groups. For his advice on ordinal modelling techniques, Dr Richard Williams (University of Notre Dame). For their helpful discussions and their encouragement to work my way the intricacies of the Rasch model, Dr David Andrich and Dr Barry Sheridan (University of Western Australia), Dr Svetlana Beltyukova (University of Toledo) and Prof Alan Tennant (University of Leeds).

Professor Peter Croft has commented on various stages of this work and drafts of papers emerging from it. I must also thank him for asking a variety of difficult questions that made me think about the difficult questions. Dr Milena Falcaro (London School of Hygiene and Tropical Medicine) for sharing her knowledge of Stata and MPlus; Ms Zoe Mayson for her friendship and support, as well as for passing on her invaluable IT skills (yes Zoe, you do deserve a PhD!); and Ms Rhian Hughes for sharing her knowledge of the Welsh language.

Finally, to my family and friends for their support in getting to the end of this thesis: especially to my husband Paul and to my parents, who have kept me sane and supported me throughout this project.

Context of the thesis

I undertook a BSc in Mathematics at the University of Sheffield, graduating in 2004. This generated my interested in studying medical statistics and I was then successful in obtaining a bursary from the Biological and Biotechnology Research Council to undertake an MSc in Biometry at the University of Reading. Following successful completion of this course in 2005, I was employed as Research Assistant in Biostatistics at the Arthritis Research UK Primary Care Centre at Keele University.

My main role in this post was to assist with the NorStOP study, and in particular to assist and develop analyses with other researchers. One of the pieces of work with which I was involved was centred on locomotor disability, and began to develop my interest in this area. Alongside this, I was taking a lead within the Research Centre on ordinal regression modelling.

After working for a year as a research assistant, I was offered the opportunity to undertake this PhD on a part time basis, as part of a Medical Research Council programme grant that extended the NorStOP study to its six-year follow-up. As such, the outline of the project was defined prior to my involvement. However, throughout the course of the PhD I have developed my own interests, notably in measurement theory and the use of continuous measures in epidemiology.

Table of contents

List of tables

Modelling the cross-sectional association of the PF-10 locomotor disability-specific items with socio-demographic factors and pain in the NorStOP

Derivation of an interval-level measure of locomotor disability using items from the PF-10

The generalisability the scoring mechanism for the locomotor disability measure

Psychometric properties of the locomotor disability measure

Modelling the longitudinal course of locomotor disability the NorStOP1: changes over time and their association with socio-demographic factors and pain

List of figures

Modelling the cross-sectional association of the PF-10 locomotor disability-specific items with socio-demographic factors and pain in the NorStOP

The Rasch unidimensional measurement model: concepts and theory

Derivation of an interval-level measure of locomotor disability using items from the PF-10

The generalisability the scoring mechanism for the locomotor disability measure

Psychometric properties of the locomotor disability measure

Approaches to the analysis of an interval-level measure over time

Modelling the longitudinal course of locomotor disability the NorStOP1: changes over time and their association with socio-demographic factors and pain

List of Boxes

List of Abbreviations

ACM	Adjacent Categories Model
ADL	Activities of daily living
BMI	Body Mass Index
BRHS	British Regional Heart Survey
CAS-HA	Clinical Assessment Study – Hand
CAS-K	Clinical Asssessment Study – Knee
CAT	Computerised Adaptive Testing
CFA	Confirmatory factor analysis
CFI	Comparative Fit Index
CI	Confidence Interval
CRM	Continuation Ratio Model
CVD	Cardiovascular disease
df	Degress of freedom
DIF	Differential Item Functioning
ELSA	English Longitudinal Study of Ageing
EU	European Union
GEE	Generalised estimating equations
GP	General Practitioner
HADS	Hospital Anxiety and Depression Scale
ICF	International Classification of Functioning Disability and Health
ICIDH	International Classification of Impairments, Disabilities and Handicaps
IQR	Inter-quartile Range
IRT	Item response theory
KAP	Keele Assessment of Participation
KGPRP	Keele General Practice Research Partnership
kgs	Kilograms
KNEST	Keele Knee Pain Cohort Study
LEFS	Lower Extremity Functional Scale

MANOVA	Multivariate analysis of variance
MCID	Minimal Clinically Important Difference
MIC	Minimal Important Change
MLRM	Multinomial Logistic Regression Model
NHSSTS	National Health Service Strategic Tracing Service
NorStOP	North Staffordshire Osteoarthritis Project
OR	Odds Ratio
PCM	Partial Credit Model
PF-10	Physical Functioning subscale of the SF-36
(P)POM	(Partial) Proportional Odds Model
PROMIS	Patient-Reported Outcomes Measurement Information System
PSI	Person separation index
REC	Research Ethics Committee
RMSEA	Root Mean Square Error of Approximation
RSM	Rating scale model
SD	Standard Deviation
SDD	Smallest detectable difference
SEM	Standard error of measurement
SF-12	Short-Form 12-item health questionnaire
SF-36	Short-Form 36-item health questionnaire
SIP	Sickness Impact Profile
SLRM	Stereotype Logistic Regression Model
SPPB	Short Physical Performance Battery
TLI	Tuker-Lewis Index
UK	United Kingdom
US(A)	United States (of America)
WHO	World Health Organisation
WHS	Welsh Health Survey
WRMR	Weighted Root Mean Square Residual

Publications and presentations arising from this thesis

Peer-reviewed publications

Muller S, Thomas E, Peat G. Derivation and testing of an interval-level score for measuring locomotor disability in epidemiological studies of middle and old age. Quality of Life Research. 2009; 18: 1341-1355

Mottram S, Peat G, Thomas E, Wilkie R, Croft P. Patterns of pain and mobility limitation in older people: cross-sectional findings from a population survey of 18 497 adults aged 50 years and over. Quality of Life Research 2008; 17: 529-539

Oral Presentations

Muller S, Thomas E, Peat G. The use of simple summated ratings scales in health research. RSS PHCSG, London 23 February 2010

Muller S, Thomas E, Peat G. Measuring locomotor disability: A new scoring mechanism for some familiar items. UK Rasch User Group Meeting, Cambridge 25 January 2010 (invited talk)

Muller S, Thomas E, Peat G. Locomotor disability in older people: exploring the influence of pain. KARMA, Aberdeen 24-25 September 2009

Muller S, Thomas E, Peat G. Measuring locomotor disability: A new scoring mechanism for some familiar items. Society for Social Medicine, Newcastle 9-11 September 2009

Mottram S, Thomas E, Peat G. Measuring locomotor disability in later life: do we need gender specific scores? International Conference on Outcomes Measurement, Bethesda 11-13 September 2008

Mottram S. Ordinal regression models: applications in epidemiology. RSS South West local meeting, Exeter 5 June 2008 (invited talk)

Mottram S, Thomas E, Peat G. The use of simple summated rating scale in health research. YSM 2008, Newport 18-20 March 2008

Mottram S, Peat G, Thomas E, Wilkie R, Croft P. The use of ordinal logistic regression models: an example from a study of mobility limitation. YSM 2007, Bristol 11-12 April 2007

Mottram S, Peat G, Thomas E, Wilkie R, Croft P. The use of ordinal logistic regression models: an example from a study of mobility limitation. RSS PHCSG, London 27 March 2007 (invited talk)

Mottram S, Peat G, Thomas E, Wilkie R, Croft P. The use of a partial proportional odds model with an ordinal response: an example from a study of mobility limitation. SSPC/SAPC/RSS PHCSG Joint Meeting 2006, Edinburgh 16-17 November 2006

Poster Presentations

Mottram S, Peat G, Thomas E, Wilkie R, Croft P. Severe Mobility limitation in middle and older age: prevalence and association with age, gender and lower limb pain in 16,007 community-dwelling adults. SAPC 2006, Keele 12-14 July 2006

1 Introduction

This thesis is concerned with the measurement of locomotor disability in population surveys of middle- and old-aged adults. For the purposes of this thesis, middle-age is defined as 50 to 64 years and old-age 65 years and over. This is in accordance with the work of Melzer et al (2005) in the English Longitudinal Study of Ageing.

1.1 What is locomotor disability?

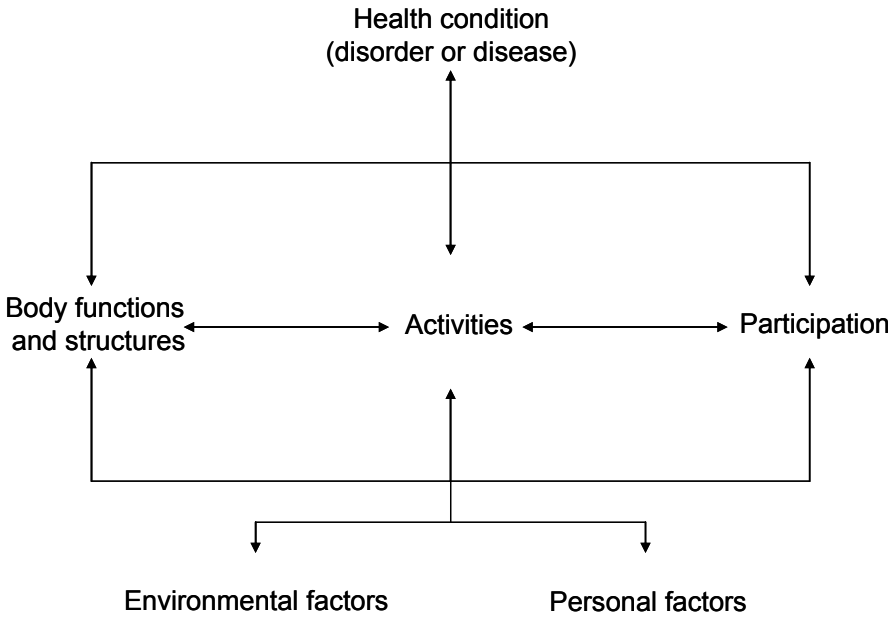
There is no agreed upon definition of locomotor disability or what it encompasses in the literature. Indeed, there is little consensus even on what to call this problem. Several terms have been used to describe difficulties with locomotion, some of which are given in Box1.1. Each could be said to have a slightly different meaning.

Box 1.1 Synonyms and related terms for locomotor disability
Gait limitations (Coppin et al 2006)
Locomotor disability (Ebrahim et al 2000)
Locomotor activity limitation (Adamson et al 2004)
Lower extremity disability (Wolinsky et al 2007)
Mobility deficit (Kokhar et al 2001)
Mobility difficulty (Leveille et al 2007)
Mobility disability (Weiss et al 2007)
Mobility impairment (Nordstrom et al 2007)
Mobility limitation (Koster et al 2007)
Mobility restrictions (Sakari-Rantala et al 2002)

Differences in terminology might be expected to arise from different and changing underlying conceptual frameworks of disability and function. For example, Martin et al (1988) considered the International Classification of Impairments, Disabilities and Handicaps (ICIDH) (World Health Organisation 1980) as a model for their national study of disability in Great Britain. This led them to use the term “locomotor disability”, which has continued in common usage, notably in British government publications, such as the Health Survey for England (Bajekal et al 2003) and in large regional (Adamson et al 2003) and national (Ebrahim et al 2000) surveys.

Other authors may have based their terminology around the successor to the ICIDH, the International Classification of Functioning, Disability and Health (ICF) (World Health Organisation 2002). The ICF is the current framework of the World Health Organisation (WHO), set out to facilitate the description of health and health-related states (WHO 2002). Under the ICF framework, "disability" is an umbrella term to denote negative function on three levels. The first level, "body function and structure", refers to disability as an impairment. An example of an impairment might be joint space narrowing due to osteoarthritis in a particular joint. At the second level, an activity is described as "*the execution of a task or action by an individual*". Hence a person experiences an activity limitation if, for example, they cannot walk 400 yards. The third level concerns "participation", which is defined as "*involvement in a life situation*", regardless of impairments or activity limitations. An example of a participation restriction might be the inability to go out to socialise with friends. Figure 1.1 shows the ICF model.

Figure 1.1 International Classification of Functioning Disability and Health (ICF) (WHO 2002)



There are further models, such as that of Nagi (Nagi 1965), from which some of these synonyms for locomotor disability could have arisen. They may also reflect idiosyncratic decisions by individual researchers to choose specific terms, which may or may not have been intended to have a specific meaning. For example, Odding et al (1995) described locomotor disability as "*the difficulty people experience when carrying out the basic activities of daily living related to the lower*

limbs", whilst Ahacic et al (2003) described mobility as "*the ability to walk and go up and down stairs*". The recent Patient-Reported Outcomes Measurement System (PROMIS) initiative refers to a domain of mobility, defining this as "lower extremity" physical functioning (Bruce et al 2009).

The ICF consists of chapters, or domains, for the different areas of life in which an individual may experience disability, one of which is mobility. This ICF chapter covers a wide variety of activities and participation domains, including "*changing and maintaining body position*", "*carrying moving and handling objects*", "*walking and moving*" and "*moving around using transportation*", and is a much broader concept than is encompassed by much of the previous research in this field (see Chapter 3).

Given the current popularity of the ICF, this thesis will adopt the term 'disability', to cover both activity limitations and participation restrictions, in keeping with the framework. However, given the broad scope of the term 'mobility' under the ICF framework, adoption of this term to reflect the lower limb function of interest in this thesis, and in much other work in this field, appears misleading. Therefore, the term 'locomotor' will be used to refer to the activity and participation domains of interest, and the term 'locomotor disability' will be used to reflect reported difficulties in these activity and participation domains. Further details will be given in Chapter 3 as to the exact activities that are considered to be key to studying this type of disability and what is meant by having difficulty with them.

Hence, in this thesis, locomotor disability will be defined as '***the difficulty a person experiences in moving from place to place using the lower limbs***'.

1.2 Consequences of locomotor disability

Locomotor disability can have many consequences. This section will consider the impact on the individual, in terms of their health, quality of life and social life as well as the effect of locomotor disability on society as a whole.

1.2.1 Individual perspective

1.2.1.1 Health impacts

Locomotor disability is frequently the first disability to arise and can occur much earlier than other problems, often in middle-age (Blazer et al 2006, Chaves et al 2000, Melzer et al 2005). It has been hypothesised that there is a hierarchy to disability, and evidence suggests that locomotor disability (Dunlop et al 1997), and in particular disability in walking (Weiss et al 2007), is at the foot of this hierarchy. Shumway-Cook et al (2005) remarked that locomotion is central to many other activities, and this could explain this early and central role of locomotor disability in the hierarchy. Weight is added to this argument by the empirical hierarchy developed by Dunlop et al (1997), where disability in walking was found to develop first, followed by disability in bathing, transferring, dressing, toileting and feeding.

Many authors point to locomotor disability as a marker for general disability, functional decline (Ahacic et al 2003, Bohannon et al 2004) and dependence on others (Fried et al 2000, Fried & Guralnik 1997). Indeed there is evidence to suggest that not only is locomotor disability associated with morbidity (Gill et al 2006, Peat et al 2006a), but also with higher levels of mortality (Gill et al 2006, Guralnik et al 2001, Mendes de Leon et al 2006, Wannamethee et al 2005). In their study of older Americans, Gill et al (2006) found that death was almost invariably preceded by self-reported locomotor disability and in a sample of New Yorkers aged 65 years and over, Khokhar et al (2001) showed that locomotor disability without disability in activities of daily living (ADL) was a predictor of mortality.

As well as these effects on physical health, locomotor disability has been associated with higher levels of affective disorders such as anxiety and depression (Iezzoni et al 2001, Shumway-Cook et al 2005). This could be as a result of reduced quality of life and social participation, issues that will be discussed in the next section.

1.2.1.2 Quality of life and social impacts

Several studies have shown an association between locomotor disability and reduced quality of life (Äijänseppä et al 2005, Chaves et al 2000). There are several reasons why this may be. Gill et al (2006) and Iezzoni et al (2001) state that locomotor disability can lead to social isolation, which has been shown to be associated with mortality and poor mental health (Seeman 1996).

Locomotion is a key factor in determining whether older people can continue to live independently (Ahacic et al 2000) and locomotor disability can lead to a dependence on other people, not only for locomotor tasks, but also for ADL and other activities. In turn, this can lead to a need for care from family and friends (Avlund et al 2003), and potentially from the state. In the United Kingdom (UK) it is estimated that 78% of older people with locomotor disability are helped by their spouse or another live-in relative, whilst 6% rely on social services or paid help (Evandrou 2005a).

It should also be considered that some 5.6% of people in the UK aged 65 years and over provide at least 20 hours per week of unpaid care (Young et al 2005). It seems inevitable that some of these people will have locomotor disability and that this could impair their ability in their role as a carer. Therefore, it would be valuable to society as a whole to enable these people to maintain their role as a carer despite having locomotor disability.

Furthermore, there are monetary costs involved for those with locomotor disability. These are difficult to assess, mainly due to the fact that people tend only to spend what they can afford and therefore studies of additional costs have been unable to determine what people would spend to aid them with their disability if they had the financial means to do so. Tibble (2005) has however produced a review of studies attempting to quantify the costs of disability. The estimates given in the review suggest a wide range of costs dependent on the methodology of the study and the type of disability considered, with most in the range £80 to £250 per person per week. These estimates are of the additional costs to persons with any disability over those with no disability. Although these costs are not specific to locomotor disability, Martin and White (1988) reported that locomotor disability accounted for the highest levels of expenditure in their study. Tibble (2005) proposed that people with disability would undoubtedly have a lower standard of living than non-disabled people with the same level of income, because of their need to spend a proportion of their income on disability-related goods and services.

1.2.2 Societal perspective

As has been discussed in the previous section, locomotor disability can reduce independence and bring about the need for social care (Fried & Guralnik 1997, Fried et al 2000). This is something that, in the UK, is often provided by the state and that society, as a whole, finances. Tibble (2005) reported that the implied level of extra costs for social assistance in those

people over pensionable age with a disability could be up to £195.95 per week. The situation will vary in other countries.

While there are no specific estimates of the added health care demand attributable to locomotor disability, due to the increased levels of both physical and psychological morbidity in those with locomotor disability, a strain is also put onto health services (Äijänseppä et al 2005, Chaves et al 2000).

These impacts and costs are important on a societal level due to the increasing proportion of elderly people in the population (Tomassini 2005) and the high levels of locomotor disability among older people (those aged 65 years and over and increasing with increasing age) (Figure 5.10: Evandrou 2005b). Guralnik et al (1996) emphasised the need to consider disability, of which locomotor disability is the most common form (Martin et al 1988), as a public health outcome. Although locomotor disability is more common at older ages, it affects a substantial proportion of those in middle age who have not yet reached statutory retirement (Iezzoni et al 2001, Melzer et al 2005). This led these authors to suggest that locomotor disability should be considered a wider public health issue, not just one which affects the oldest age-groups, in order to retain those in middle-age in the workforce.

1.3 Thesis aims

1.3.1 Rationale and hypothesis

1.3.1.1 Hypothesis

The hypothesis underlying this thesis is that because locomotor disability lies on a continuum, it is most appropriately and usefully measured on an interval-level scale. The following section gives the rationale behind this hypothesis.

1.3.1.2 Rationale

Many health conditions have generally been regarded as being present or absent. However, this does not follow the natural occurrence of diseases. Geoffrey Rose proposed that the vast majority of diseases exist on a continuum,

"*...the idea of a sharp distinction between health and disease is a medical artefact for which nature, if consulted, provides no support...disease is nearly always a quantitative rather than a categorical or qualitative phenomenon, and hence it has no natural definitions*". (Rose 1992; pg 6-8).

More recently, Thomas (2007) has made a similar point, arguing for a disability continuum, with no sharp distinction between function and disability,

"*Disability, objectively considered, is a continuum along which all individuals in the general population are ranged (though not evenly), from the Olympic athlete at one end to the comatose patient at the other....Between the extreme groups lies a third large group whose disabled/non-disabled status is a matter of judgement.*" (Thomas 2007).

In certain circumstances, such as the allocation of treatment (Rose 1992; pg 9) or social support (Thomas 2007), it is undoubtedly necessary to draw a distinction between the diseased and healthy or disabled and non-disabled. However, this approach is not necessarily suitable for public health, where it can give the impression that those who do not fall into the 'disease' group and so do not quality for treatment (in the case of disease), or assistance or support (in the case of disability) are assumed not to be 'at risk' of some adverse outcome. Rose (1992; pg 11-12) clearly illustrates this phenomenon with an example of blood pressure levels and risk of heart attack or stroke in the following 18 years. Although a decision to treat may be made at a diastolic blood pressure of 100 mgHg, those with blood pressure below this level are not immune from a heart attack or stroke over the next 18 years.

As will be described in Chapter 3, locomotor disability is often regarded as a dichotomous phenomenon, with people defined as 'disabled' or 'non-disabled'. Although this may be necessary for some health planning purposes, it does not allow a full epidemiological investigation of the phenomenon of locomotor disability. In longitudinal studies, a dichotomy does not take into account the magnitude of change in underlying disability. This may mean that clinically relevant changes in the level of disability are missed because the threshold for disability is not crossed, or that due to measurement error, people with no clinically relevant change in their level of disability may be seen to cross the threshold (Sakari-Rantala et al 2002).

A continuous measure of disability would allow the magnitudes of differences in disability to be quantified, and hence also the rate of change in disability to be compared across groups. Should the factors associated with the onset of, or recovery from, locomotor disability using a

binary definition be different from those factors associated with high rates of change in disability, then the choice of whether to 'define' or to 'measure' locomotor disability becomes of paramount importance.

In addition to the discussion of the continuum of disease, Geoffrey Rose's seminal works of the 1980s (Rose 1982, 1985) considered the difference between the health of populations and the health of individuals. Rose argued that the way to make real differences to the health of a population was to consider the factors affecting the health of that population, rather than the factors affecting the health of individuals within that population,

"*...what distinguished two groups is nothing to do with the characteristics of individuals, it is rather a shift of the whole distribution - a mass influence acting on the population as a whole.*"

(Rose 1985).

This argument is relevant not just to diseases, as Rose discussed, but also to disability. Locomotor disability is the most common disability in the community (Martin et al 1988) and is particularly common in older people (Bajekal et al 2003). As the population ages then, locomotor disability will become a major public health concern. The availability of a continuum on which to measure disability would allow a more thorough and natural investigation into the extent of disability in the population, its effects on those with disability, and its potential causes. More importantly, from the public health perspective, the measurement of locomotor disability on a continuum would allow the distribution of disability across populations to be compared, to ascertain the population-level characteristics that differ between those groups with low levels of disability and those with higher levels. This would ultimately aid the development of public health interventions to reduce the impact of locomotor disability on the individual and society.

1.3.2 Specific objectives

The specific objectives identified in order to assess the hypothesis underlying this thesis are:

1. to assess the current state of measurement of locomotor disability and related concepts in the literature, and give an overview of the basic epidemiology in this field.

2. to assess the suitability of current approaches to the measurement of locomotor disability: in particular, to assess the measurement properties of the Short Form-36 (SF-36) Physical Functioning subscale (PF-10) for this purpose.

3. to explore the use of ordinal regression models to analyse data from single items of the PF-10 as an intermediate approach to analysis, between single-item binary definitions and multiple-item interval-level measurement.

4. to use Rasch analysis techniques to derive an interval-level measure of locomotor disability suitable for use in a general population using selected items from the PF-10 and to test the psychometric properties of this measure in a general population.

5. to illustrate the potential of the new measure of locomotor disability in analyses of the epidemiology of locomotor disability in a longitudinal setting, over and above what is possible using a binary definition of disability.

1.4 Thesis overview

Chapter 2 What is measurement?

Different levels of data are encountered in epidemiological studies. This chapter describes these levels and the relationships between them. It concludes by defining the term 'measurement' as it will be used throughout this thesis.

Chapter 3 The measurement of locomotor disability and its occurrence in middle- and old-age: a systematic search and overview

The current state of the measurement of locomotor disability and its epidemiology, both descriptive and analytic, at the start of this PhD project are summarised and evaluated.

Chapter 4 The North Staffordshire Osteoarthritis Project (NorStOP)

The North Staffordshire Osteoarthritis Project (NorStOP), provided the majority of data for the empirical analyses presented in this thesis. The data from this study are assessed for their suitability to address the objectives of this thesis. Potential strengths and weaknesses of the NorStOP dataset in this regard are highlighted and discussed.

Chapter 5 The required properties of measurement instruments

Properties that have previously been stated as being necessary for health measurement instruments to possess are described and summarised as they will be used in this thesis.

Chapter 6 *The suitability of the SF-36 Physical Functioning subscale (PF-10) as a measure of locomotor disability*

The PF-10 has been suggested as a potential interval-level measure of locomotor disability. This chapter assesses the suitability of the PF-10 as such a measure and considers the measurement properties of selected items from the scale.

Chapter 7 *Approaches to modelling ordinal outcome data*

An overview of the possible approaches to modelling of ordinal dependent variables in regression models is presented, covering the specific uses of and problems with each.

Chapter 8 *Modelling the cross-sectional association of the PF-10 locomotor disability-specific items with socio-demographic factors and pain in the NorStOP*

Individual items from the PF-10 are used to assess the cross-sectional association of locomotor disability with socio-demographic factors and pain in the NorStOP.

Chapter 9 *The Rasch unidimensional measurement model: concepts and theory*

The background to and theory of the Rasch unidimensional measurement model are described and the criteria for its further use in later chapters set out.

Chapter 10 *Derivation of an interval-level measure locomotor disability using items from the PF-10*

An interval-level measure of locomotor disability is derived from the five locomotor disability-specific items from the PF-10.

Chapter 11 *The generalisability of the scoring mechanism for the locomotor disability measure*

The generalisability of the scoring mechanism for the interval-level measure of locomotor disability derived in Chapter 10 is tested in three external datasets: the Keele Knee Pain Cohort Study, the Welsh Health Survey and a Dutch cohort.

Chapter 12 Psychometric properties of the locomotor disability measure

The psychometric properties of the measure of locomotor disability developed in Chapter 10 are tested in the NorStOP dataset. Properties are assessed against those laid out in Chapter 5.

Chapter 13 Approaches to the analysis of an interval-level measure over time

Possible methods of analysing interval-level data in a large scale population epidemiology study are described.

Chapter 14 Modelling the longitudinal course of locomotor disability the NorStOP1: changes over time and their association with socio-demographic factors and pain

Possible uses of the new locomotor disability measure in a longitudinal epidemiology study are illustrated in the NorStOP1. In particular, changes in the score over time and their association with changes in pain status are considered.

Chapter 15 Discussion and conclusions

The principal findings of this thesis are reviewed before the major decisions taken in the course of this PhD project are discussed, along with their potential impact of the findings. The chapter concludes with a discussion of the implications of this thesis for future research in this field.

2 What is measurement?

A key aim of this thesis is to consider the appropriate way in which to measure locomotor disability in a general population. In order to do this, a clear definition of the concept of measurement is needed. This chapter describes the different types or 'levels' of data that are found in health research, the relationships between them, and the suitability of each for the purpose of measurement.

2.1 Levels of measurement

Quantitative data can be provided on several levels: nominal, ordinal, interval and ratio (Wharrad 2004). Nominal- and ordinal-level data are categorical, whilst interval- and ratio-level data are continuous.

2.1.1 Categorical data

Categorical data, as the name suggests, are data that consist of categories. There are two types of categorical data: nominal and ordinal. The distinction between these two types of categorical data is discussed below.

2.1.1.1 Nominal data

Nominal data consist of two or more categories with no order. The simplest form of nominal data is binary or dichotomous. For example, gender is male or female. An example with more categories might be the type of arthritis a patient has: osteoarthritis, rheumatoid arthritis, psoriatic arthritis, other.

2.1.1.2 Ordinal data

Ordinal data consist of ordered categories. These ordered categories occur in two principal forms: grouped continuous and qualitative.

Grouped continuous data arise from the splitting of an underlying continuum. An example of this would be body mass index (BMI) classification. BMI is a continuous variable, calculated from

height and weight, and can be grouped according to whether a person is underweight (<18kg/m^2), normal (18.5 to 24.98 kg/m^2), overweight (25 to 29.9 kg/m^2) or obese (≥30kg/m^2) (WHO 2006).

Qualitative categories result from subjectively assessing to which category a response belongs. It is assumed that an assessor, for example a survey respondent, has assimilated information from an indeterminate number of sources in order to provide a judgement. An example of this type of data would be when people are asked to rate their health as “excellent”, “very good”, “good”, “fair” or “poor” in the Short Form-12 health survey (SF-12) (Ware et al 1996). This type of data is often referred to as 'assessed' or 'judged'. Table 2.1 gives some more examples of different types of ordinal data that may occur in epidemiological studies.

Any numbers applied to the categories of ordinal data represent only the rank order of the categories and not their relative magnitudes (Merbitz et al 1989). For example, in the PF-10 item shown in Table 2.1, “no, not limited at all” is given the value 3, “yes, limited a little” is given the value 2 and “yes, limited a lot” is given the value 1. This does not imply that “yes, limited a little” is equivalent to “no, not limited at all” minus “yes, limited a lot”.

2.1.2 Continuous data

2.1.2.1 Interval-level data

Interval-level data, such as temperature, are continuous and points are represented by equidistant numbers on the scale are an equal distance apart. Hence the mathematical operations of addition and subtraction can be performed. For example, the difference between 3°C and 8°C is 5°C, as is the difference between 41°C and 46°C. It does not matter what point along the scale is considered, a difference of five points represents a difference of 5°C. Interval-level data do not have a meaningful zero point, i.e. a temperature of 0°C does not mean no temperature, and so it does not make sense to say that 20°C is twice as warm as 10°C. Hence, the functions of addition and subtraction are appropriate, but multiplication and division are not.

2.1.2.2 Ratio-level data

Ratio-level data are similar to interval-level data, except that there is a meaningful zero point. Hence, addition, subtraction, multiplication and division are all appropriate mathematic operations to be used with ratio-level data. Consider for example, weight, measured in kilograms (kgs). A

weight of 0kgs implies and means no weight, so it makes sense to say that 20kgs is twice the weight of 10kgs.

Table 2.1 Examples of ordinal data: grouped continuous and qualitative categories

Grouped continuous		
Construct of interest	Original level of measurement	Ordinal response variable
Blood pressure (National Institute for Health & Clinical Excellence 2006)	Continuous ≥ 0 (mmHg)	≤140: normal >140 & ≤160: hypertension >160: drug treatment
Number of consultations in a 12-month period (Rowlands & Moser 2002)	0,1,2,3,…	0: None 1-2: Below average 3-4: Average 5+: Above average
Qualitative		
Construct of interest	Example of a question	Response options and coding in questionnaire
Arthritis Impact Measurement Scales 2 (Meenan et al 1992)	During the past month how would you describe the hand pain you usually had?	Severe - 1 Moderate - 2 Mild - 3 Very mild - 4 None - 5
Illness Perceptions Questionnaire – Revised (Moss-Morris et al 2002)	Please indicate your agreement or disagreement with the following possible cause of ill health: Stress or worry.	Strongly disagree - 1 Disagree - 2 Neither agree not disagree - 3 Agree - 4 Strongly agree - 5
PF-10 (Ware & Sherbourne 1992)	Does your health now limit you in walking more than a mile? If so, how much?	No, not limited at all - 3 Yes, limited a little - 2 Yes, limited a lot - 1

Measures of health status do not usually require a meaningful zero point to make sense. For example, if measuring mood, there is no point when there is no mood (Wright & Linacre 1989). In some cases though, these zero points occur naturally, for example the number of stairs a person can climb without assistance could be zero.

2.2 Ordinal-level scales versus interval-level measures

Ordinal- and interval-level data are often considered interchangeably. This may be because ordinal data are misinterpreted as being on an interval-level (Merbitz et al 1989), especially when the ordinal scale is very long; say seven or more individual categories, and so it is assumed that the distinction between data types is inconsequential.

The problem with this assumption is that in order to calculate a change in score, or indeed the difference in score between two groups, the data need to be suitable to be used in arithmetic operations. As discussed above, the arithmetic operations of addition and subtraction only make sense when conducted on interval- or ratio-level data (Merbitz et al 1989). For multiplication and division, data must be on a ratio-level.

Figure 2.1 illustrates why treating ordinal scales as interval- or ratio-level is inappropriate. Assuming that the data are ordinal (bottom of the ruler), Person A starts at a score of five points at Time 1, and increases to a score of seven points at Time 2. This is a change of two points on the ordinal scale, but a change of over three points on the interval-level measure (top of the ruler). Person B also increases their score by two points on the ordinal scale (13 points to 15 points) from Time 1 to Time 2, but on the ordinal scale their score increases by less than two points. This can also be reversed, whereby equal changes on an interval-level measure do not equate to equal changes on an ordinal scale. Hence changes on the interval- and ordinal-level scales are not necessarily equivalent at different points along such scales.

There are readily available non-parametric procedures with which to analyse ordinal-level data in a suitable manner (Tennant et al 2004). However, the objective of many studies in health research is to consider changes in a score over time, or between groups. For this type of analysis, as already discussed, interval-level measurement is required in order to use mathematical and statistical operations appropriately.

A major barrier to the creation of interval-level measures in health research seems to have been the apparent ‘success’ of the use of ordinal scales in mathematical and statistical procedures. This is due to the naturally occurring ogival shape of the relationship between ordinal scores and interval-level measures (Wright & Linacre 1989). Towards the centre of the ogive, where many of the data are located, the relationship is approximately linear (Figure 2.2), but it fails towards the extremes of the distribution (Wright & Linacre 1989).

Figure 2.1 Change scores on interval- and ordinal-level scales: differences over time or between groups

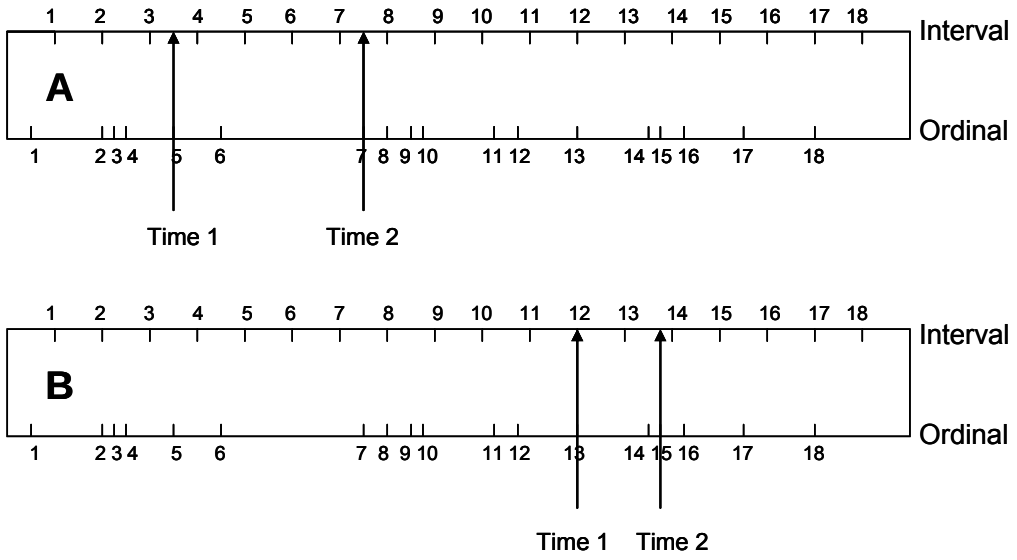
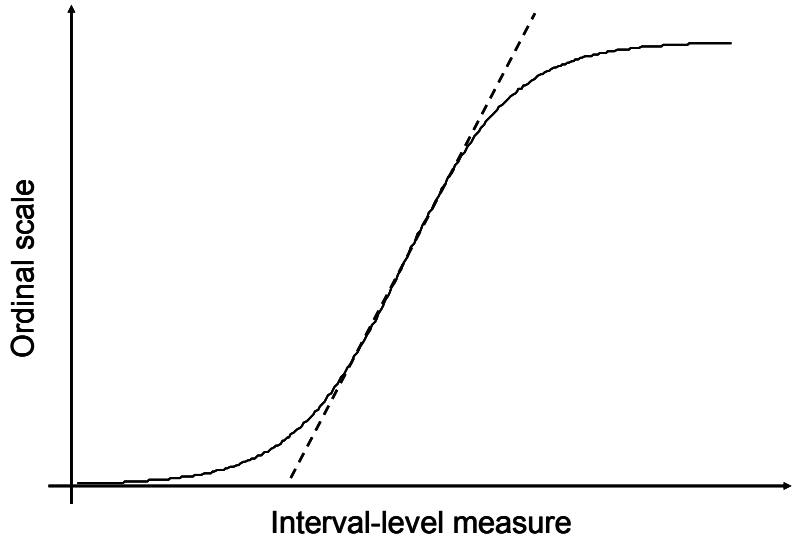


Figure 2.2 The relationship between ordinal-level scales and interval-level measures: an ogive



Tennant (2007) used simulated data to compare change scores for ordinal- and interval-level data, and showed that for 14% of individuals substantially different change scores were achieved. This difference was not systematic, with some change scores for the ordinal-level scale being smaller than for the interval-level measure and others larger. Although this difference may balance out across large numbers of people, there is no way of knowing this is the case, and if scores are used at the individual level, as they often are in health measurement, these differences could be misleading and result in suboptimal treatment decisions. Tennant related this difference in change scores between ordinal-level scales and interval-level measures to their divergence at the extremes, of their distributions, as shown in Figure 2.2.

With the exception of the study of Tenant (2007), there is little empirical evidence assessing the potential impact on study conclusions of treating ordinal-level scales as though they were interval-level measures. This could be due to the difficulties in conducting this type of study, which as discussed by Tennant (2007), include the attempted linearization of the extreme categories of an ordinal-level scale, and the necessity of using simulated rather than 'real' data in order to be reassured that the ordinal-level scale can be made linear.

In addition to the study of Tennant (2007), several authors have provided theoretical arguments against the use of ordinal-level scales as though they were interval-level (Merbitz et al 1989, Pae 1999, Svensson 2001, Tennant et al 2004). For example, Merbitz et al (1989) stated that the use of ordinal-level scales in this manner might distort the evidence around the effectiveness of treatments, whilst Tennant et al (2004) highlighted the potential consequence of distorted effect sizes. It is reasonable to extrapolate potential distortions of calculated statistics outside the realms of studies of treatment effect and into more general epidemiological settings, where changes over time and sizes of effects are still very much of interest.

2.3 Summary

This chapter has outlined the levels of data found in health research. These levels are important to understand, both in terms of the direct interpretation of the data and the correct handling of data and statistical analyses.

There are strong theoretical arguments against the analysis of ordinal-level scales as though they were interval-level measures, and indeed empirical evidence that demonstrates the potential errors that could be induced through this approach is beginning to appear. It is therefore prudent to

ensure that data are suitable for the intended purpose and that methods of analysis are suitable for the data in question.

For the purpose of this thesis, measurement will be defined as '***the use of an interval- or ratio-level scale to assess an underlying construct'***. In order to distinguish between levels of data in this thesis, the terms 'measure' and 'measurement' will only be used in reference to interval- or ratio-level data. Data on nominal- or ordinal-level scales, even if these scales are long, will not be referred to as measures.

The next chapter considers the literature relating to the measurement of locomotor disability in middle- and old-age at the start of this PhD project. In particular, the review considers the range of self-report instruments previously used to assess locomotor disability, and how these instruments compare to objective, performance-based measures. This is followed by a summary of the basic epidemiology of locomotor disability in community-based studies of older adults.

3 The measurement of locomotor disability and its occurrence in middle- and old-age: a systematic search and overview

3.1 Introduction

This thesis is concerned with the measurement of locomotor disability in epidemiological studies. This chapter therefore provides a background to the measurement of locomotor disability (Section 3.3.1), and summarises its epidemiology (Sections 3.3.2 to 3.3.5), both descriptive and analytic, at the start of this PhD project.

The specific objectives of this review were:

- to describe the range of self-report instruments for locomotor disability (and related concepts) used in previously published epidemiological studies involving adults aged 50 years and over: specifically instrument content, item phrasing, and response options;
- to compare the use of self-report instruments for locomotor disability (and related concepts) with objective, performance-based assessments;
- to describe the estimates of prevalence rates of locomotor disability (and related concepts) in adults aged 50 years and over;
- to describe the rates of onset and cumulative incidence of locomotor disability (and related concepts) in adults aged 50 years and over;
- to describe the rates of recovery from locomotor disability (and related concepts) in adults aged 50 years and over;
- to describe the factors associated with locomotor disability (and related concepts) in adults aged 50 years and over.

3.2 Methods

3.2.1 Search Strategy

A systematic search was carried out to collect the available evidence at the start of this project. Searches were carried out in the databases Medline, EMBASE and DH-DATA from their

inception to the start of the review (1950-2007, 1974-2007 and 1983-2007 respectively), using the strategy outlined in Appendix A. This strategy was based around a comprehensive set of synonyms for 'locomotor disability' and was compiled using the Medical Subject Headings assigned to known key papers in Medline.

3.2.2 Eligibility criteria

Specific inclusion and exclusion criteria for the review were specified.

Inclusion criteria

- Observational studies
- Studies using community-based samples
- Studies assessing locomotor disability via self-report methods
- Studies comparing self-reported locomotor disability with objectively assessed performance
- Studies including participants aged 50 years and over

Exclusion criteria

- Clinical trials and other randomised studies
- Studies based exclusively on samples taken from non-community-based settings, such as care homes or those consulting a medical professional
- Studies based exclusively on assessed performance-based tasks
- Studies of populations exclusively under the age of 50 years
- Studies of disability where the assessment of locomotor disability was not reported separately from other forms of disability

3.2.3 Selection of articles

The titles and abstracts of retrieved papers were assessed for suitability for inclusion according to the criteria outlined above. Where there was doubt as to the suitability of the study, the full paper was obtained and a decision about inclusion was made based on this. All selection of articles was undertaken by a single reviewer (SM).

3.2.4 Data extraction and synthesis

Data were extracted from the selected articles by a single reviewer (SM). Due to the wide variety of formats of data found in these papers, and the range of data required for this review, a formal data extraction form was not used. Instead, information relevant to each section of the review was extracted separately in a narrative manner.

All papers were considered for inclusion in the assessment of previously used self-report instruments described in Section 3.3.1.1. Studies comparing objectively assessed performance to self-reported locomotor disability were considered for inclusion in Section 3.3.1.2.

All studies identified in the original selection process were eligible for consideration in describing the prevalence of locomotor disability in Section 3.3.2. For the description of the onset and cumulative incidence of, and recovery from locomotor disability in Sections 3.3.3 and 3.3.4, by definition, only longitudinal studies were included.

Due to the large number of papers regarding the factors associated with the onset of and recovery from locomotor disability, this review considered only longitudinal studies of these associations (Section 3.3.5). This had the advantage, not only of narrowing down the volume of studies to be considered an area that has received considerable research attention, but also of ensuring that associations were temporal, one of the requirements for causality (Hill 1965). Factors associated with the onset of and recovery from locomotor disability were considered separately because it is not necessarily the case that the removal of a causal factor will result in recovery, although in some cases this may be true.

When considering the rates of the prevalence and onset of and recovery from locomotor disability, studies from the UK were considered first, and then international studies were considered separately. Due to the heterogeneous nature of the data extracted in this review, all syntheses were carried out in a narrative manner.

3.2.5 Additional searches

The search strategy used in this review is similar to that applied by Stuck et al (1999) in their review of the epidemiology surrounding functional status decline. However, the difficulty in defining locomotor disability and the differing terminology used in the literature mean that it is impossible to be certain that all relevant papers were included in this review.

To improve the comprehensiveness of the review, searches were carried out in the United Nations Disability Statistics Database (United Nations 2007), UK Data Archive (University of Essex 2007) and the Economic and Social Research Council's Question Bank (University of Surrey 2007), in addition to the searches of the major health care databases described above. To improve comprehensiveness further, the reference lists of all eligible papers were searched for studies that may have been missed in the database search.

3.3 Results

3.3.1 Assessment of locomotor disability using self-report instruments

The definition of locomotor disability is problematic, as discussed in Chapter 1. Even more challenging is its assessment, which can be carried out using two broad approaches: self-report by the participant, or the observation of functional performance. This section reviews the self-report instruments that have been used to assess locomotor disability and their comparability to observed functional performance.

3.3.1.1 Self-reported assessment

There were no instances in the literature of studies assessing locomotor disability that used true interval-level measures of self-reported disability, as defined in Chapter 2. This section therefore focuses on how locomotor disability has previously been assessed and defined.

Self-reported assessments of locomotor disability have been used in many studies. However, there is little consensus as to how locomotor disability should be assessed in the general population. Some studies have used recognised tools for assessing disability or functional limitation that include items on locomotion, such as the Rosow and Breslau (1966) scale used by Guralnik et al (1993), the Health Assessment Questionnaire (Fries et al 1982) used by Odding et al (1995) and the PF-10 (Ware & Sherbourne 1992) used by Peat et al (2006a). Many studies though have used a selection of items from different tools or devised new items.

There are three distinct areas to an instrument that assesses locomotor disability: tasks or actions included, framing and phrasing of the question(s), and specifying response options. Each is considered in turn in this section.

Tasks or actions included

Different studies have asked about a multitude of different activities, connected in varying degrees to the concept of locomotion (Table 3.1). This is a common problem in the field of disability measurement (Swanson et al 2003). The activities covered in the studies in this review ranged from those obviously related to locomotion, such as walking, which was covered by most studies, to the slightly more ambiguous, such as cutting toenails (Kriegsman et al 1997).

More than half of studies asked about walking in some way. However, the distances that were considered varied substantially between studies; from ‘across a small room’ (Guralnik et al 2001) to more than a mile (Bohannon et al 2004). Several distances in-between these were measured in a number of different units; metres, yards, blocks.

After walking, the most common activity considered was ascending/descending stairs. Questions were often worded slightly differently to each other, with some insisting that the task be completed without help (Leveille et al 2000) or without resting (Iezzoni et al 2001), whilst others did not specify (Ahacic et al 2003).

Activities considered less often in relation to locomotor disability included, but were by no means limited to, bending, straightening, transferring to and from a bed, chair or car and balancing. The number of falls in the previous year was also considered in some studies (Adamson et al 2003).

Table 3.1 Locomotor tasks considered in studies included in this review

Locomotor task	Studies considering this task
Walking	
More than half a mile[a]	Bohannon et al (2004), Melzer & Parahyba (2004), Peat et al (2006a), Sainio et al (2006)
Half mile[a]	Avlund et al (1996), Chaves et al (2000), Clark et al (1998a and 1998b), Fried et al (2000), Guralnik et al (1993, 2001), LaCroix et al (1993), Leveille et al (2000, 2002), Mendes de Leon et al (2006), Merrill et al (1997), Nordstrom et al (2007), Peat et al (2006a), Salive et al (1994), Visser et al (1998)
Between 100 yards and half a mile [a,b]	Adamson et al (2004), Äijänseppä et al (2005), Chang et al (2004), Ebrahim et al (2000), Fried et al (2000), Gill et al (2006), Hirani & Malbut (2002), Iberg et al (2001), Iezzoni et al (2001), Launer et al (1994), Leveille et al (2007), Melzer et al (2005), Sainio et al (2006), van den Brink et al (2004), Wannamethee et al (2005)
100 yards[b]	Ahacic et al (2000, 2003), Bohannon et al (2004), Clark et al (1998a), Melzer & Parahyba (2004), Peat et al (2006a), van den Brink et al (2004)
Inside the home	Evandrou (2005b), Guralnik et al (2001), Iberg et al (2001), Khokhar et al (2001), Launer et al (1994), Mendes de Leon et al (2006), Sainio et al (2006), Sakari-Rantala et al (2002),
Outside the home	Adamson et al (2004), Äijänseppä et al (2005), Avlund et al (2002), Ebrahim et al (2000), Evandrou (2005b), Khokhar et al (2001), Odding et al (1995), Office for National Statistics (2003), Sakari-Rantala et al (2002), van den Brink et al (2004), Wannamethee et al (2005)
No specified distance or location	Adamson et al (2003), Heikkinen et al (1983), Martin et al (1988), Odding et al (1995)
Stair climbing (various wordings and numbers of stairs)	Adamson et al (2003 and 2004), Ahacic et al (2000 and 2003), Äijänseppä et al (2005), Avlund et al (1996 and 2002), Bohannon et al (2004), Chaves et al (2000), Clark et al (1998a and 1998b), Ebrahim et al (2000), Evandrou (2005b), Fried et al (2000), Gill et al (2006), Guralnik et al (1993), Heikkinen et al (1983), Iberg et al (2001), Iezzoni et al (2001), Khokhar et al (2001), Kriegsman et al (1997), LaCroix et al (1993), Launer et al (1994), Leveille et al (2000, 2002, 2007), Martin et al (1988), Mendes de Leon et al (2006), Merrill et al (1997), Nordstrom et al (2007), Odding et al (1995), Office for National Statistics (2003), Peat et al (2006a), Sainio et al (2006), Sakari-Rantala et al (2002), Salive et al (1994), Steel et al (2004), van den Brink et al (2004), Visser et al (1998), Wannamethee et al (2005)
Bending, straightening, stooping, crouching or kneeling	Adamson et al (2003 and 2004), Bohannon et al (2004), Ebrahim et al (2000), Iberg et al (2001), Launer et al (1994), Martin et al (1988), Odding et al (1995), Peat et al (2006a), Sakari-Rantala et al (2002), Steel et al (2004), Wannamethee et al (2005)
Moving and/or carrying objects	Äijänseppä et al (2005), Bohannon et al (2004), Clark et al (1998b), Fried et al (2000), Iberg et al (2001), Launer et al (1994), Peat et al (2006a), Sainio et al (2006), Steel et al (2004)
Transferring to/from and/or rising from various places, for example a chair, a bed, a car	Chaves et al (2000), Fried et al (2000), Iberg et al (2001), Khokhar et al (2001), Launer et al (1994), Leveille et al (2002), Odding et al (1995), Sainio et al (2006), Steel et al (2004)

Other activities (including housework and self-care activities)	Bohannon et al (2004), Fried et al (2000), Iberg et al (2001), Leveille et al (2002), Melzer & Parahyba (2004), Merrill et al (1997), Peat et al (2006a)
Keep balance	Adamson et al (2004), Ebrahim et al (2004), Martin et al (1988), Wannamethee et al (2005)
Falls	Adamson et al (2003 and 2004), Martin et al (1988),
Running (various distances)	Ahacic et al (2000, 2003), Sainio et al (2006)
Use of public and/or private transport	Khokhar et al (2001), Kriegsman et al (1997), Sainio et al (2006)
Cutting toe nails	Kreigsman et al (1997)
Standing	Iezzoni et al (2001)
Use of mobility aid	Iezzoni et al (2001)

[a]Half a mile is approximately equivalent to 805 metres, this is also regarded as "several blocks"; [b]100 yards is approximately equivalent to 91 metres, this is also regarded a "one block"

Framing and phrasing the questions

There are many ways in which questions about locomotor disability can be asked. They can relate to capacity, i.e. what a person is capable of doing, or performance, i.e. what a person actually does. Most studies identified in this review used capacity-based items.

There were two major ways of framing questions about locomotor disability: those that did and those that did not make reference to health. Those questions that referenced health states generally took the form, "*had difficulty carrying out…on their own as a result of a long term health or medical problem or due to old age*" (Adamson et al 2004) or "*does your health limit you in…*" (Bohannon et al 2004). Questions not making reference to health tended to be more straightforward. For example, "*can you…?*" (Ahacic et al 2003) or "*how difficult is it for you to…*" (Clark et al 1998a).

The different framings used may gain different responses. In their study of life roles disability in older adults, Dubuc et al (2004) found higher levels of reported disability in a variety of activities when the item made no attribution to health as the origin of the disability than when this attribution was made. These authors attributed this difference in response to other factors that might influence disability, for example the physical and social environment. This seems particularly relevant to the items in the study of Dubuc et al (2004), which were concerned with disabilities in life roles. These authors found that the difference between attribution and non-attribution items was greater in items where a significant effect of the environment could be envisaged, for example "*travel out of town*" and "*visit friends and family*". It is not clear how health attribution and the interaction of the

individual with the environment might translate into the field of locomotor disability, where many items do not directly involve such interactions.

Related to these issues is context. If locomotor disability is assessed in a survey predominantly about health then people may assume disability needs to be health-related, regardless of question wording. However, if the questionnaire is broader, for example The General Household Survey (Office for National Statistics 2003), then disability may be reported in that wider context (Thomas 2007).

Most papers do not quote questionnaire phraseology directly and therefore the framing used for questions in a particular study is often unclear. However, where quotes are given, the timeframe is often ambiguous. For example, Chaves et al (2000) used a question phrased, "*For health or physical reasons, do you have difficulty climbing up 10 steps?*". Here there was no reference to a time period. In the development of the PF-10, Ware et al (2000) stated that they considered functional limitations to be chronic and therefore did not specify a timeframe. Conceptually this is difficult to follow, as it implies that recovery is not possible, an assumption that may not be true. It is unclear though whether other studies where timeframes are non-specific also expected disability to be chronic. Further studies have given vague timeframes: "*currently*" (Ebrahim et al 2000), "*nowadays*" (Saino et al 2006), "*normally*" (Melzer & Parahyba 2004). Few studies have described exact timeframes, for example "*one week*" (Chang et al 2004). The lack of a specific time frame could potentially make answering difficult for respondents, particularly if they experience health problems that are intermittent or vary in severity over time.

As described above, many studies used the word 'difficulty' in questions, for example 'how much difficulty do you have…?'. This word does not have standard meaning across researchers and participants (Thomas 2007). In the WHO's Disability Assessment Schedule II Manual, "*...having difficulty with an activity means increased effort, discomfort or pain, slowness or changes in the way the person does the activity*" (WHO 2000; pg 16). It is not clear however, whether all respondents interpret the word "difficulty" in the same way. In a qualitative study of elderly, American women, Porter (2007) demonstrated that there was "*difficulty rating difficulty*", as participants would often prefer to use another word, or found that their problem varied over time. In another study, Gregory & Fried (2003) ascertained that older adults were able to cite reasons for reporting difficulty with a task. The most common of these reasons was having had to modify the way in which they carried out a task. Other reasons for reporting difficulty included a slower rate of

task completion, cutting back on the number of times they performed a task and experiencing pain whilst performing a task. One would imagine that similar words such as “limitation” or “disability” can also be interpreted in a number of ways and that the same reasons for these interpretations would apply as with difficulty. However, there is currently no empirical evidence to support or refute this assertion.

Many activities included in the locomotion items used in questionnaires, even those that are seemingly straightforward, such as walking a specific distance, have been criticised as difficult for respondents to conceptualise. For example, in her qualitative research, Mallinson (2002) found that some people do not know if they can walk a mile, as they would never attempt to do it, preferring instead to go by car. Others have little sense of how far a particular distance actually is: they work on local landmarks, not measured distances. Furthermore, Mallinson (2002) described how some people may be able to walk a particular distance on the flat but not uphill. In their study, Melzer and Parahyba (2004) considered access to the necessary ‘resources’ to do a task. Their study was carried out in Brazil and included areas of ‘shanty towns’, they therefore felt that many people would not have access to stairs, so they asked only about walking limitations, avoiding the introduction of a socioeconomic bias.

These criticisms of the items in self-reported assessments may be valid, but they are largely unavoidable in questionnaires or structured interviews. Culturally appropriate activities can be chosen and the use of specific terms in the phrasing of self-report items is clearly preferable to non-specific terms, for example ‘walk half a mile’ as opposed to ‘walk to the supermarket’, but many of the issues raised in this section cannot be fully overcome.

Specifying response options

Having established how questions are framed and phrased in locomotor disability assessment instruments, a related issue is that of the possible response options and how to transform these options into useable scores. Again this varied greatly between studies.

Consider first what response options might be available. Many studies in this review allowed a respondent to answer “Yes” or “No” to an item, so they could be categorised as able or unable to perform a particular activity. Alternatively, more options were given. Clark et al (1998a) gave five possible response options to the question of how difficult it is to perform a task: “not at all”, “a little”, “somewhat”, “very”, “don’t do”. Other studies (Bohannon et al 2004, Peat et al 2006a) used

questions from the PF-10 (Ware & Sherbourne 1992), which allows three options: “yes, limited a lot”, “yes, limited a little”, “no, not limited at all”. Table 3.2 shows the types of response options used in the studies included in this review.

Melzer et al (2004) considered there to be a continuum underlying the categories of disability and carried out a study in an American population to assess whether the thresholds between the option categories changed across different population subgroups. They found that males and females had similar thresholds, as did those with different levels of education. However, whites, older people and those with lower incomes tended to have lower thresholds than non-whites, younger people and those with higher incomes respectively, i.e. they reported disability at a higher level of functioning, as assessed by the MOBility-related Limitation Index (Lan et al 2002). Melzer et al (2004) also suggested that different environments might give rise to different responses at similar levels of health. For example, those in very hot or cold climates may find walking outdoors more difficult than those in more temperate climates.

Table 3.2 Response options used in studies included in this review

Response options	Studies using these response options
Dichotomous	Adamson et al (2003 and 2004), Ahacic et al (2000 and 2003), Avlund et al (1996 and 2002), Chang et al (2004), Chaves et al (2000), Clark et al (1998b), Ebrahim et al (2000), Evandrou (2005b), Gill et al (2006), Guralnik et al (1993), Iberg et al (2001), Khokhar et al (2001), LaCroix et al (1993), Launer et al (1994), Leveille et al (2000), Martin et al (1988), Melzer & Parahyba (2004), Melzer et al (2005), Mendes de Leon et al (2006), Merrill et al (1997), Nordstrom et al (2007), Office for National Statistics (2003), Sakari-Rantala et al (2002), Salive et al (1994), Steel et al (2004), Wannamethee et al (2005)
Polytomous	
3 categories	Bohannon et al (2004), Clark et al (1998a), Fried et al (2000), Hirani & Malbut (2002), Kriegsman et al (1997), Peat et al (2006a), van den Brink et al (2004)
4 categories	Äijänseppä et al (2005), Odding et al (1995), Iezzoni et al (2001), Sainio et al (2006), Visser et al (1998)
5 or more categories	Clark et al (1998a), Leveille et al (2007),

Thomas (2007) discussed the arbitrary nature of assigning a ‘disability label’ to an individual. One person, for example an athlete, may clearly have no problematic locomotor disability, whereas a bedridden person clearly has problematic disability. However, there will be many people in between these two extremes, and when a disability score is assigned to a person, it is necessary,

in certain circumstances, to be able to define a threshold at which that person is considered 'disabled'. There is no consensus on where this threshold should be and whether it varies between population subgroups, for example males and females.

Many studies used one or more questions with dichotomous responses in order to derive an overall binary variable for locomotor disability, for example difficulty on one or more items defined a person as having disability (Table 3.3). Other studies summed the number of items endorsed by the respondent to get an index (for example, from 0 to 3) and then either dichotomised (Mendes de Leon et al 2006), or calculated a mean (Odding et al 1995), in order to define a binary disability variable. Other authors used similar methods based on polytomous items, whilst some used more complex grading systems.

3.3.1.2 Self-report versus observed functional performance

The previous section discussed the advantages and disadvantages of self-reported assessment of locomotor disability in terms of definition, phraseology and response scoring. This section will go on to consider the comparability of these self-reported assessments with performance in locomotor tasks.

Observed functional performance assessments of locomotor disability involve physically measuring the abilities of a research participant or patient. Tasks are generally quite simple: timing a walk over a set distance or to stand from a chair, or assessing ability to climb a flight of stairs.

Observed assessments could be considered more useful than self-reports as they are more objective. For example, they are less susceptible to social acceptability bias and are less influenced by culture, language and education level (Merrill et al 1997). They are also less ambiguous for participants in that, for example people do not have to gauge distances and whether they can walk that far. This judgement aspect of self-reported measures is particularly problematic in older people, where they may only walk short distances in their daily lives and so in answering a question about a longer distance in self-report assessments they are speculating on both the distance and their capabilities (Chang et al 2004). Given a set course, it is possible to test a person's ability to walk a particular distance, even if, for practical reasons, this is not the same distance that might be asked about in a self-report measure and therefore may not be completely comparable. For example, the speed at which a participant walks 10 metres, when they know that

they only have to walk that distance, might not be reflective of their normal walking speed over longer distances.

Table 3.3 Methods of scoring/defining locomotor disability used in studies included in this review

Method	Studies using this method
Positive response to individual dichotomous item or items defines 'disability'	Evandrou (2005b), Fried et al (2000), Gill et al (2006), Hirani & Malbut (2002), Iberg et al (2001), Melzer et al (2005), Office for National Statistics (2003), Sakari-Rantala et al (2002)
Positive response to one of a set of one or more dichotomous items defines 'disability'	Adamson et al (2003, 2004), Äijänseppä et al (2005), Avlund et al (1996), Chaves et al (2000), Clark et al (1998b), Ebrahim et al (2000), Guralnik et al (1993), Khokhar et al (2001), LaCroix et al (1993), Launer et al (1994), Leveille et al (2000), Merrill et al (1997), Nordstrom et al (2007), Salive et al (1994), Steel et al (2004), Wannamethee et al (2005)
Count of number of dichotomous items on which difficulty is reported	Ahacic et al (2003), Avlund et al (2002), Mendes de Leon et al (2006)
Polytomous items dichotomised at specified level and	
dichotomised polytomous items used as individual binary responses	Sainio et al (2006)
one or more items with value above the specified level defined 'disability'	Kriegsman et al (1997), van den Brink et al (2004), Visser et al (1998)
count the of number of items with values above the specified level	Clark et al (1998a)
(Transformed) sum of polytomous item responses	Bohannon et al (2004), Peat et al (2006a), Visser et al (1998)
Mean of polytomous item responses with predefined threshold imposed on mean defined 'disability'	Odding et al (1995)
Individual polytomous item responses used to represent levels of disability	Hirani & Malbut (2002)
Complex grading of disability based on series of binary or polytomous item responses	Guralnik et al (2001), Iezzoni et al (2001), Leveille et al (2007), Martin et al (1988), Melzer & Parahyba (2004)

Tests are usually undertaken indoors on a flat, even surface and the environment is standardised for all participants. This allows comparisons to be made between participants, and measures the physiological component of locomotion (Sainio et al 2007). However, what is not accounted for in these tests is the participant's daily experience of locomotor activities, which is likely to be quite different, involving environmental, cultural and attitudinal components, as well as

physiological aspects (Sainio et al 2007). For example, walking outside might involve the weather, an uneven pavement, or a slope. Therefore, a person might be able to walk at the required speed in a laboratory test, but in their own environment, be unable to walk to the local shop or bus stop. This “real life” situation, possibly best classified as participation, may be more important to people, and it can only be reasonably captured using self-report instruments. However, as has already been discussed, there are problems with self-report: people do not always fully comprehend the questions or they may never attempt the activity. Response options to express these situations are rarely provided.

Several studies have considered the relationship between self-reported and observed functional performance assessments of locomotor disability. Chang et al (2004) found that 80% of people aged 75 to 85 years, reporting that they could walk a quarter of a mile (402 metres) could actually walk 400 metres.

Sainio et al (2007) compared self-reported ability to climb a flight of stairs with observed performance in climbing and descending two steps in a population of Finns aged 45 years and over. Overall agreement, as assessed by the Kappa statistic was 0.58 (95% CI 0.54, 0.61), which can be considered moderate agreement (Altman 1991; pg 404). Using the observed performance as the standard, the authors found that 10% of people over-reported and 34% under-reported their ability to climb stairs. Levels of over-reporting were higher in females than in males. However, as discussed by these authors, the two definitions of disability were not equivalent in that the observed task involved fewer stairs and also involved descending.

Sayers et al (2004) used three questions from the Walking Performance Interview (Guralnik et al 2003) as their self-reported assessment and compared this to a timed 400-metre walk. They found that of those who appeared most disabled from the questionnaire, 89% were unable to complete the walk within 15 minutes and of those who appeared least disabled, only 5% were unable to complete the walk.

Bohannon et al (2004) used self-reported ability in any of three walking activities (one block, several blocks, more than a mile) to classify older women as having locomotor disability or not. They compared this to gait speed, classifying women as disabled if their walking speed was less than 1.22 metres per second over a 7.62 metre course. Agreement was poor (absolute agreement 64%, Kappa 0.152). Most discrepancies were as a result of the women reporting difficulties, but having acceptable gait speeds. These authors ascribe these discrepancies to the difference in the

distances related to the two assessments of disability and suggest that women may not feel gait speed to be important when considering their walking ability.

It has been proposed that the gap between self-reported and observed assessments could be explained by the incorporation of health expectations, health knowledge, society and culture into a person's self-report (Iberg et al 2001), whereas observed performance considers only the physiological aspects of locomotion. Regardless of the arguments surrounding the potential benefits and downfalls of self-report and observed functional performance, in large, population-based studies it is unlikely that it will be practical, in terms of time or cost, to observe locomotor performance. Therefore, self-report measures are needed in order to assess locomotor disability in these populations. This section has shown that there is a gap between the two types of assessment, but in many cases, the items compared were not equivalent. For the majority of large-scale epidemiological studies, self-report will provide a more practical method of assessing locomotor disability, as it is easier to administer and probably gives a truer reflection of the daily locomotor experience of the respondent.

3.3.2 The prevalence of lower limb locomotor disability

Vast numbers of estimates of the prevalence of locomotor disability are available in the literature. As the data used in the later stages of this thesis are primarily from the UK, prevalence estimates from this region are most pertinent and are presented first, before consideration is given to international prevalence estimates and changes in prevalence over time. The prevalence of locomotor disability will be evaluated in the population as a whole and in age and gender specific strata. Further stratification, for example by socioeconomic status, will not be considered as associations of other factors with locomotor disability will be addressed in Section 3.3.5.

3.3.2.1 UK locomotor disability prevalence estimates

Table 3.4 describes the UK studies providing prevalence estimates, which were identified by the search. The wide range of prevalence estimates for locomotor disability in the UK, from 2% in the study of Wannamethee et al (2005) to 36% in the study of Adamson et al (2004) can be seen in Table 3.5. The differences in prevalence estimates might result from slightly different age and gender compositions in the samples, or perhaps more likely, from the different definitions of locomotor disability applied.

One further study (Adamson et al 2003) was found, but as it was based solely in the west of Scotland, the sample was not representative of the UK population. The prevalence of locomotor disability in this study was 27.2%, which is similar to some of the nationally representative estimates.

Table 3.4 Studies describing the prevalence of locomotor disability in the general UK population

Reference	Study	Study type	Population	Mobility tasks
Adamson et al (2004)	British Women's Heart and Health Study	Postal survey	Representative of UK females, 60 to 79 years, n=4,286	Go up and down stairs; Bend down; Straighten up; Keep balance; Go out of the house; Walk 400 yards
Ebrahim et al (2000)	British Regional Heart Study	Postal survey	Representative of UK males, mean age 63 years, n=5,717	Go up and down stairs; Bend down; Straighten up; Keep balance; Go out of the house; Walk 400 yards
Evandrou (2005b)	General Household Survey 2001: Focus on Older People	Interview	Representative of UK population aged 65 years and over	Use stairs; Go outdoors; Walk down the road; Get around the house
Hirani & Malbut (2002)	Health Survey for England 2000	Interview	Representative of English population aged 65 years and over, n=1,677	Walk 200 metres; Climb 12 stairs
Martin et al (1988)	OPCS Surveys of Disability in Great Britain	Interview	Representative of Great British population aged 16 years and over	Items included: Walking; Falls; Balance; Stair use; Bending and straightening
Melzer et al (2005)	English Longitudinal Study of Ageing	Interview	Representative of English population aged 50 years and over, n=11,392	Walk quarter of a mile
Office for National Statistics (2003)	General Household Survey 2001	Interview	Representative of Great British population aged 65 years and over	Walk down road unaided; Stairs and steps
Steel et al (2004)	English Longitudinal Study of Ageing	Interview	Representative of English population aged 50 years and over, n=11,392	Climb stairs; Stoop/kneel/ crouch; Lift/carry; Get up from chair
Wannamethee et al (2005)	British Regional Heart Study	Postal survey	Representative of UK males aged 52 to 73 years, n=5,075	Get outdoors; Walk 400 yards; Climb stairs

Table 3.5 Prevalence of locomotor disability in the UK older adult population

Study	Locomotion task	Prevalence (%)
Adamson et al (2004)	Any one of six items (see Table 3.4)	36.5
Ebrahim et al (2000)	Any one of six items (see Table 3.4)	25.0
Hirani & Malbut (2002)	Walk 200 metres Climb 12 stairs	26.0 21.0
Martin et al (1988)	Any of thirteen items (see Table 3.4)	9.3
Office for National Statistics (2003)	Walk down road unaided Stairs and steps	14.0 10.0
Wannamethee et al (2005)	Any one of three items (see Table 3.4) All of three items (see Table 3.4)	6.5 1.9

Evandrou (2005b), Melzer et al (2005) and Steel et al (2004) provide only age or gender stratified prevalence estimates (see Table 3.6)

Table 3.6 shows the prevalence of locomotor disability in the UK older adult population, stratified by age and gender. At older ages, locomotor disability was more prevalent; this was particularly true in females. Iberg et al (2001) found evidence that males described their locomotor ability as excellent at lower levels of functioning than females, whilst Merrill et al (1997) suggested that females are more prone to reporting disability, as they are socialised to notice pain and discomfort. These factors could partly explain this gender difference. However, these possible factors do not account for the larger gap between the genders at older ages. Leveille et al (2000) conducted a seven-year study in older Americans to assess the reasons for these gender differences in prevalence. They found that the main cause of the difference was a higher rate of onset in females, with females having the onset rate of males approximately five years older. They also found the relative impacts of onset, recovery and death were different at different ages. Hence these findings could explain the larger gender gap at older ages.

Harwood (1996) suggested that prevalence estimates could be affected, not only by the assessment of locomotor disability, but also by the age and gender distribution of the population. Furthermore, it has been investigated whether the country or culture in which the study is conducted affects prevalence estimates (Iberg et al 2001, Melzer et al 2004). Hence, the next section presents international estimates of the prevalence of locomotor disability.

Table 3.6 Prevalence of locomotor disability in the UK older adult population stratified by age and gender

Study	Locomotion task	Age-group	Prevalence (%)	
			Male	Female
Adamson et al (2004)	Any one of six items (see Table 3.4)	60 to 69 years	31.4	
		70 to 79 years	42.5	
Evandrou (2005b)	Use stairs	65 to 69 years	6.0	
		70 to 74 years	7.0	
		75 to 79 years	10.0	
		80 to 84 years	16.0	
		85 years and over	24.0	
		65 years and over	22.0	29.0
Hirani & Malbut (2002)	Any one of two items (see Table 3.4)	65 to 79 years	27.0	27.0
		80 years and over	47.0	57.0
Melzer et al (2005)	Walk quarter of a mile	50 to 64 years	9.0	8.0
		65 to 79 years	17.0	20.0
		80 years and over	36.0	47.0
Steel et al (2004)[a]	Climb several flights of stairs	50 years and over	48.4	59.7
	Stoop, kneel or crouch		47.3	58.2
	Lift or carry heavy weights		28.6	55.3
	Get up from a chair after sitting for a long period		33.3	41.6

[a]Prevalence estimates given in original publication as “up to”, reflecting the highest prevalence across age bands

3.3.2.2 International locomotor disability prevalence estimates

Several studies have made international and cross-cultural comparisons of the prevalence of locomotor disability (Table 3.7). These studies showed differences, in some cases quite large, in the estimated prevalence of locomotor disability between nations and cultures. In the main, these differences could not have resulted from different activities included in the definition of locomotor disability, or from differences in question wording, as the studies were deliberately developed to account for these differences. Sakari-Rantala et al (2002) discussed how in their study populations of Jyväskylä, Finland and Glostrup, Denmark, there has previously be shown to been little difference in the prevalence of ADL disability (Avlund et al 1996) and this was attributed to the basic level of these universally necessary activities. These authors proposed that this could be the case for locomotor disability too, and the majority of their data corroborated this (Table 3.7). Similarly, Harwood (1996) discussed a 1983 report of the WHO, which found no differences in the prevalence of walking and stair climbing ability across 11 non-UK European communities (Heikkinen et al 1983). However, other studies refute this notion (Äijänseppä et al 2005, Mendes de Leon et al 2006), suggesting there is some underlying difference between populations.

Melzer et al (2004) investigated the difference in the estimated prevalence of self-reported locomotor disability in white American and Dutch populations. They found that the level of functional ability at which Americans reported disability was lower than that of the Dutch. They called this phenomenon "*response category cut-point shift*" and concluded that the differences between the two nations were not as large as the raw prevalence estimates implied, in terms of objective measures of locomotor disability. In a similar study of ethnic groups in the United States of America (USA), Iberg et al (2001) showed that whites reported excellent levels of locomotion at lower levels of functioning compared to non-whites. This was in contrast to the study of Melzer et al (2004), which showed that white Americans reported disability at lower levels of functioning than blacks. This difference could be due to a number of factors including the use of regional samples within the USA, different statistical methodology, or different definitions of locomotor disability. These cultural differences are in keeping with the hypothesis that it is more acceptable in some cultures than others to admit to or report disabilities (Sakari-Rantala et al 2002), although they do not necessarily explain where or why these differences occur.

Table 3.7 Inter-population comparisons of estimates of the prevalence of locomotor disability

Study	Activities	Population	Prevalence (%)			
Äijänseppä et al (2005)	Difficulty in any one of: move outdoors, use stairs, walk 400 metres, carry 5 kg	*Males*	*Northern Europe*	*Southern Europe*		
		80 to 84 years	52.5	40.4*		
		85 to 89 years	67.7	55.0*		
		90 years and over	83.9	80.0		
		Females				
		80 to 84 years	83.6	68.5*		
		85 to 89 years	75.6	77.0		
Andrews et al (1986, cited in Harwood 1996)	-	Adults	*Malaysia*	*Korea*	*Philippines*	*Fiji*
			15.0	15.0	29.0	42.0[a]
Melzer et al (2004)	USA: Walking quarter of a mile; The Netherlands: walking for five minutes		*USA*		*The Netherlands*	
		60 years and over	47.1		17.4*	
Mendes de Leon et al (2006)	Difficulty in any one of: walk across a small room, half a mile, walk up and down stairs	65 years and over	*Black Americans*		*White Americans*	
			18.7		15.4*	
Sakari-Rantala et al (2002)			*Jyväskylä, Finland*		*Glostrup, Denmark*	
	Walking indoors	Males	39.0		45.5	
	Walking outdoors		54.4		56.8	
	Climbing stairs		54.4		58.4	
	Walking indoors	Females	42.6		50.3	
	Walking outdoors		47.5		64.9*	
	Climbing stairs		64.4		72.7	

- Not specified; * $p<0.05$; [a]Significance of difference unknown

Whether or not international differences in prevalence estimates are real, or a product of a "response category cut-point shift" (Melzer et al 2004), the fact remains that prevalence estimates varied widely between studies, as they did when interest is restricted only to UK studies. It would have been interesting to compare international prevalence estimate to those from a UK population, as this would have allowed the UK to be viewed in a wider context, but unfortunately, no such studies were found in this comprehensive search of the literature.

3.3.2.3 Changes in the prevalence of locomotor disability over time

Prevalence estimates for locomotor disability apply only to the time at which they were calculated. It is possible for these estimates to change over time, within the population in which they were first estimated. This is interesting in relation to the compression of morbidity hypothesis (Fries 1980), which states that average life expectancy is increasing, whilst the maximum life span remains the same. Hence acute disease is being superseded by chronic illness, which is being pushed into the later years of life. This suggests that there will be an increase in disability in the population, occurring at older ages, whilst there is less disability at younger ages.

The Office for National Statistics has recently reported that whilst both overall life expectancy and disability-free life expectancy are increasing in the UK, overall life expectancy is growing at a faster rate than disability-free life expectancy (Office for National Statistics 2008). This means that the average length of time spent with a disability is increasing, and should logically lead to an increase in the prevalence of disabilities, including locomotor disability, in the population. Furthermore, in a study across European Union (EU) member states, Jagger et al (2007a) demonstrated that whilst average life expectancy was similar across countries, the average healthy life expectancy was higher in those states that had been members of the EU for longer. This suggests an association between levels of economic development and the potential scale of the problem of disability in older persons.

This section considers studies examining these changes in prevalence over time using repeated cross-sectional samples. Studies are only included if the eligibility criteria and mode of administration remained consistent over waves of the survey. The individuals in the study are however expected to differ, at least in part, between the waves.

In the Health Survey for England 2001, Bajekal et al (2003) showed a significant increase in the proportion of males and in the population aged 16 years and over that experienced locomotor

disability between 1995 and 2000/2001 (Table 3.8). Unfortunately these data were not reported by age-group and hence data specifically regarding older adults could not be derived, although these data do provide an estimate of the increased burden of locomotor disability in the population over this period. However, as shown above, the prevalence of locomotor disability is strongly age related and so an increase in overall prevalence may not be indicative of a higher prevalence of disability throughout the population, merely of a change in the age structure of that population.

Table 3.8 Change in prevalence of locomotor disability in the English population aged 16 years and over (Bajekal et al 2003)

	Prevalence (%)		Relative change[a] (95% confidence interval)
	1995	2000/2001	
Male	9.2	10.8	17 (8, 27)
Female	13.0	13.8	6 (1, 14)
Overall	11.3	12.5	11 (2, 19)

[a]Relative change: $(Prevalence_{2000\text{-}2001} - Prevalence_{1995})/Prevalence_{1995}$

Evandrou et al (2005b) also showed an increase in the prevalence of locomotor disability in a British population aged 65 years and over from 1980 to 2001, with 14% unable to walk outdoors alone and 10% unable to climb stairs in 2001, compared to 12% and 8% respectively in 1980. As the sample was restricted to those aged 65 years and over, the potential for this difference to have been due to a change in the age structure of the population was reduced but not eliminated.

Although this increase in prevalence over time could be due to an increase in the proportion of people in the oldest age-groups, there could be other explanations. For example, it is possible that the changes seen in the studies detailed above are real changes in the age-specific level of locomotor disability in the population. Alternatively this could be a 'cohort effect', when membership of a certain group or cohort influences an outcome and affects prevalence estimates over repeated cross-sectional studies. Those aged 65 years and over in 1980 were born in or before 1915, whilst those aged 65 years and over in 2001 were born in or before 1936. Hence their life experiences are different. It could be that those born prior to 1915 are, as a group, stoical, having lived through World Wars I and II. These people might then always report fewer ailments at any age or in any time period, than the younger cohort. Locomotor disability would then appear more prevalent in

2001 than in 1980. Another possibility is a so-called 'period effect', whereby something about the time of reporting changes a person's response. In this case, it could be more acceptable to report physical problems in 2001 than in 1980 and so everyone is more likely to report locomotor disability, regardless of the cohort to which they belong, resulting in a higher prevalence estimate. Before trying to untangle these effects though, it is worth considering a large-scale Swedish study.

Ahacic et al (2000) reported on a sample of approximately 6,000 people that was representative of the Swedish population aged 18 to 75 years in 1968 and 1991. Ahacic et al (2000) found that the prevalence of locomotor disability decreased over the 23-year period in those aged 50 years and over. This difference was more noticeable in the oldest old. An example of this reduction is a change in the prevalence of disability from 41% to 24% in stair-climbing and from 37% to 20% in walking 100 metres in those aged 70 to 75 years. On average, mobility limitations in the Swedish population were shown to be "postponed" by approximately 10 years over the study period, i.e. a prevalence estimate for a particular age-group in 1991 was approximately equal to that of those 10 years younger in 1968, providing evidence in support of the compression of morbidity hypothesis.

Clearly the results of this Swedish study are in contrast to the UK studies. This could be a real difference between the nations, or a period or cohort effect as discussed above. It is difficult to tell which of these reasons is most likely, as these effects can never be fully disentangled (Menard 2002), and it is likely that one or more of them is occurring concurrently. However, the results of the Swedish study (Ahacic et al 2000) are in agreement with the now widely accepted (Mor 2005, Jagger et al 2007b) compression of morbidity hypothesis, meaning that locomotor disability seems to be being pushed to a later point in life and hence the prevalence of the disability is decreasing over time. It is also likely though that the ageing population (Tomassini 2005) will offset this effect as the numbers of people in the oldest age-groups increase. The results from the Health Survey England are not age stratified and it is likely that the ageing population and consequent increase in the numbers of the oldest old are responsible for the apparent increase in the prevalence of locomotor disability. The acceptance of the compression of morbidity hypothesis may seem to lessen the need to study locomotor disability. However, with the ever increasing average life expectancy (Fries 1980) and the large number of middle-aged people that will reach old-age in the coming decades, the absolute size of this problem is set to dramatically increase. A decrease in the

prevalence of locomotor disability does not negate the public health issues surrounding this disability.

3.3.3 Onset and cumulative incidence of locomotor disability

Onset and cumulative incidence of locomotor disability are both terms that refer to the new occurrence of disability over a period of time. However, they have slightly different definitions: onset is the proportion of people with no disability at the start of the follow-up period, who are disabled at the end; cumulative incidence is the proportion of people with no disability at the start of the follow-up period who develop disability at some point during the follow-up. There is a distinction made here because estimates of the rate of onset do not necessarily capture all incident cases of disability arising during the study period. If there are high levels of recovery among incident cases, onset will underestimate cumulative incidence (Gill et al 2006). However, the majority of studies in this review considered onset rather than incidence, as studies of the former are much less resource intensive and are easier to conduct.

This section of the review will focus first on UK studies and then examine international studies of onset and incidence. Differences in onset rates between different populations will also be described.

3.3.3.1 UK locomotor disability onset and cumulative incidence estimates

Only one UK study of onset was found in this review. Wannamethee et al (2005) used a postal survey to study a representative sample of 5,075 British males aged 52 to 73 years. Over a four-year period they found an onset of locomotor disability, defined as difficulty in going up or down stairs, walking 400 yards or going out of the house, of 10.5%.

Table 3.9 shows the age stratified estimates of the onset of locomotor disability from this study. There is little obvious trend, although onset rates are higher in those aged 70 years and over than in younger males.

Table 3.9 Onset of locomotor disability in British males, by age (Wannamethee et al 2005)

Age-group at baseline (years)	Onset (%)
52 to 55	8.8
55 to 59	10.3
60 to 64	8.4
65 to 69	9.9
70 and over	14.3[a]

[a]Test for trend p=0.03

3.3.3.2 International locomotor disability onset and cumulative incidence estimates

Table 3.10 shows the background to the international studies of onset to be considered in this section.

Tables 3.11 and 3.12 summarise the findings of these papers, both overall and stratified by gender respectively. There is variation in the estimates of onset, from 6% in the study of Clark et al (1998a) to 26% in the study of Clark et al (1998b). However, the age structure of these populations differs (Table 3.10) and, as shown by the studies of Wannamethee et al (2005) (Table 3.9) and Guralnik et al (1993) (Table 3.13), the onset of locomotor disability is higher at older ages.

One study considered the cumulative incidence of locomotor disability in a population of Americans aged 70 years and older (Gill et al 2006). They found that there were frequent transitions to disability, with a five-year cumulative incidence of 60.6% in stair-climbing and 81.6% in walking quarter of a mile. This suggests much higher levels of incidence than suggested by most of the onset studies, although again this could be an issue of definition.

Table 3.10 International studies describing the onset of locomotor disability

Study	Study type	Population	Locomotion tasks
Avlund et al (2002)	Interview	75 years, n=510, general population, Denmark and Finland	Mob-H Scale (includes getting outdoors; Walking outdoors in nice weather; Walking on stairs)
Chaves et al (2000)	Interview	Female, 70 to 80 years, n=436, Medicare population, USA	Walk 800 metres; Climb 10 steps; Transfer to/from a car/bus
Clark et al (1998a)	Postal survey	51 to 61 years, n=5,017, general population, USA	Walk one block; Walk several blocks; Climb one flight of stairs without resting
Clark et al (1998b)	Interview	70 years and over, n=2,857, general population, USA	Walk several blocks; Climb one flight of stairs; Pull/push heavy objects; Carry 10 pounds
Guralnik et al (1993)	Interview	65 years and over, n=6,978, general population, USA	Walk up and down stairs; Walking half a mile without help
Khokhar et al (2001)	Interview	65 years and over, n=586, Medicare population, USA	Walking (indoors, outdoors); Climbing stairs; Rising from chair; Crossing streets; Disembarking train/bus
Leveille et al (2000)	Interview	65 years and over, n=10,263, general population, USA	Walk up and down stairs; Walking half a mile without help
Sakari-Rantala et al (2002)	Interview	75 years, n=519, general population, Denmark and Finland	Transfer to/from bed/chair; Walking (indoors, outdoors); Climbing stairs
Visser et al (1998)	Interview	65 years and over, n=4,809, general population, USA	Walk 800 metres; Walking up 10 steps

Table 3.11 International estimates of the onset of locomotor disability

Study	Locomotion task	Period of study	Onset (%)
Chaves et al (2000)	Any one of three items (Table 3.10)	18 months	23.9
Clark et al (1998a)	Any one of three items (Table 3.10)	2 years	6.0
Clark et al (1998b)	Any one of four items (Table 3.10)	2 years	25.5
Khokhar et al (2001)	Any one of six items (Table 3.10)	2 years	14.2

Table 3.12 International estimates of the onset of locomotor disability, by gender

Study	Locomotion task	Period of study	Age-group at baseline	Onset (%)	
				Male	Female
Avlund et al (2002)	Mob-H Scale (Table 3.10)	5 years	75 years	10.0	9.0
Visser et al (1998)	Any two items (Table 3.10)	3 years	65 to 100 years	14.8	20.3
	Walking 800 metres			11.8	15.0
	Walking up 10 steps			7.7	11.7

Visser et al (1998) (Table 3.12), Leveille et al (2000) (exact values not provided in original publication) and Sakari-Rantala et al (2002) (Table 3.13) all reported a higher level of onset of locomotor disability in females than in males. There is little evidence to suggest reasons for a gender difference in onset, although it has been proposed that health expectations are higher in females (Iberg et al 2001). Leveille et al (2000) showed that once disabled, females are more likely than males to remain disabled; hence a higher rate on onset will be detected because rates of recovery are lower. The same may not be seen when considering cumulative incidence.

Guralnik et al (2001) investigated the four-year course of locomotor disability in an American population aged 65 years and over. They found that in those aged 65 to 74 years, approximately 60% of severe disability was progressive rather than catastrophic in onset: that is 60% of participants acquired moderate disability before the onset of severe disability. This percentage increased with age, particularly in males. In those aged 85 years and over, more than 80% of disability was progressive in onset.

Table 3.13 presents the levels of onset of locomotor disability found in different localities. These data show that, as with prevalence, the rate of onset can vary across populations even when definitions are the same. Reasons for this may be similar to those for prevalence: differences in perception, health expectations and cultural acceptability.

3.3.4 Recovery from locomotor disability

Recovery from locomotor disability has different meanings in different studies, in that it may mean less disability or return to previous abilities. It is usually defined however, as not reporting disability or difficulty in a locomotor task where such disability or difficulty was previously reported.

In this section, recovery will first be considered in UK studies and then in international populations.

Table 3.13 Onset of locomotor disability in different localities within studies

Study	Population group	Locomotion task	Onset (%)		
			East Boston	*Iowa*	*New Haven*
Guralnik et al (1993)[a]	*Male*	Any one of two items (Table 3.10)			
	65 to 74 years		34.0	27.0	44.0
	75 to 84 years		60.0	50.0	67.0
	85 years and over		79.0	77.0	93.0
	Female				
	65 to 74 years		36.0	27.0	50.0
	75 to 84 years		58.0	51.0	71.0
	85 years and over		70.0	84.0	91.0
			Jyväskylä		*Glostrup*
Sakari-Rantala et al (2002)[b]	*Male*	Walk indoors	28.0		33.3
		Walk outdoors	47.2		45.1
		Climb stairs	44.4		44.7
	Female	Walk indoors	39.8		42.9
		Walk outdoors	59.5		59.3
		Climb stairs	49.1		58.5

[a]Development of locomotor disability over a four-year period; [b]Development of locomotor disability over a five-year period

3.3.4.1 UK locomotor disability recovery estimates

Wannamethee et al (2005) reported a recovery rate, i.e. loss of all disability, of 27.1% over four years in a representative sample of British males aged 52 to 73 years at baseline. Table 3.14 shows that recovery rates were lower in younger males, although this was not statistically significant.

3.3.4.2 International locomotor disability recovery estimates

Table 3.15 shows the reported levels of recovery from locomotor disability in three international studies. The lowest level of recovery, at 9%, was seen in the American population of Kokhar et al (2001) aged 65 years and over, with the highest level of recovery, at 56%, in the study of Clark et al (1998a) in those aged 51 to 61 years. This suggests that recovery is more likely in middle-aged than in old-aged persons.

In contrast to the findings of Wannamethee et al (2005) in their British study, Leveille et al (2000) described a lower recovery rate with older age (exact data not provided in original

publication). There is little other evidence on the age-related trend in recovery and it is not possible to ascertain whether this is a real effect or simply an artefact of these data.

Table 3.14 Recovery from locomotor disability by age in British Males (Wannamethee et al 2005)

Age-group at baseline (years)	Four-year recovery rate (%) [a]
52 to 55	22.5
55 to 59	24.0
60 to 64	27.0
65 to 69	28.0
70 and over	28.9

[a]Test for trend p=0.26

Table 3.15 International estimates of recovery from locomotor disability

Study	Population[b]	Period of study	Recovery (%)
Clark et al (1998a)	51-61 years, USA, n=1,420	2 years	56.0
Clark et al (1998b)	70 years and over, n=1,871, USA	2 years	25.5
Khokhar et al (2001)[a]	65 years and over, USA, n=480	2 years	8.8

[a]Excluding those with locomotor disability and ADL disability at baseline; [b]Further details of these studies can be found in Table 3.10

3.3.5 Factors associated with locomotor disability

3.3.5.1 Factors associated with the onset of locomotor disability

This section will consider the factors associated with the onset of locomotor disability. Groups of factors will be considered: socio-demographic, lifestyle, and health conditions and impairments. Some studies consider baseline characteristics in those without locomotor disability and assess disability status at follow-up, other consider changes in characteristics over the follow-up period.

Socio-demographic factors

Several studies have shown an association between female gender and the onset of locomotor disability. Clark et al (1998a, 1998b) showed that over a two-year follow-up period females aged 51 to 61 years and 70 years and over respectively, were approximately one and a half times more likely than males to become disabled. In their seven-year follow-up study, Leveille

et al (2000) showed that in three American communities and across age-groups, females were more likely to be disabled than males and that this gap widened with age. Gill et al (2006) described how female gender was also associated with the cumulative incidence of locomotor disability.

Older age has been shown to be associated with a higher onset of locomotor disability. Clark et al (1998b) showed that those aged over 85 years were more than twice as likely to develop locomotor disability over a two-year period than 70 to 74 year olds. In the Established Populations of Epidemiological Studies of the Elderly cohorts, Guralnik et al (1993, 2001) showed a significant age-related increase in the four-year onset of locomotor disability in adults aged 65 years and over. Gill et al (2006) showed a similar pattern of age-related cumulative incidence in their five-year follow-up study.

In a multi-national study in Europe, van den Brink et al (2004) showed that men aged 70 years and over who became widowed during a five-year follow-up period, were more likely to develop locomotor disability than those who remained married. Widowers living alone experienced a lower level of onset than widowers living with others. This however, argue the authors, could be a case of reverse causality: locomotor disability, or a precursor to it, could have caused these widowed men to be living with other people at the start of the study.

Many studies considered the association between socioeconomic status and locomotor disability. In a nationally representative cohort of British men, Ebrahim et al (2000) and Wannamethee et al (2005) showed a higher rate of onset in those with a lower social class, as derived from job title. This finding has been replicated in studies in a Swedish population (Ahacic et al 2003), although there is some evidence in this Swedish study that class differences are diminishing over time. Other definitions of socioeconomic status have shown similar patterns, with those in lower socioeconomic groups having a higher likelihood of the onset of locomotor disability (fewer years of education (Clark et al 1998a, Sainio et al 2007); lower income (Guralnik et al 1993); blue collar workers (Ahacic et al 2000); living in a disadvantaged neighbourhood (Nordstrom et al 2007)) than those in higher groups. Saino et al (2007) showed that socioeconomic differences in locomotor disability were largely due to the higher prevalence of obesity, chronic disease and a history of strenuous work in lower socioeconomic groups. Such lifestyle and related health factors will be considered further below.

Avlund et al (2002) showed that those aged 75 years living in sheltered housing were seven times more likely to develop locomotor disability by the age of 80 years than those living independently in the community. However, no other studies have corroborated this and, despite the prospective nature of the study, it could be that people entering sheltered housing were predisposed to locomotor disability.

Clark et al (1998a, 1998b) have investigated the possible association between the onset of locomotor disability and non-white race in the USA. However, they failed to show any significant association.

What is clear from these studies is that the effect of socioeconomic conditions from early- and middle-life, as well as later life, influence locomotor disability onset into older age.

Lifestyle factors

Four main lifestyle factors have been investigated with respect to their association with locomotor disability: bodyweight, tobacco smoking, alcohol consumption and physical activity. Each of these will be considered in turn. BMI has been associated with the onset of locomotor disability in a range of population groups (Bannerman et al 2002, Clark et al 1998a and 1998b, Ebrahim et al 2000, LaCroix et al 1993, Launer et al 1994, Mendes de Leon et al 2006, Wannamethee et al 2005), although the association was not always the same in these groups. For example, Mendes de Leon et al (2006) showed that the association was different in black and white Americans with black people being less disabled at higher levels of BMI than whites. Launer et al (1994) suggested that there might be several mechanisms by which obesity is associated with locomotor disability. These include the direct association of increased wear and tear on joints, and the indirect mechanism whereby high BMI results in morbidity, which in turn causes locomotor disability. Substantial weight changes (gain or loss) have been shown to be significantly associated with the onset of locomotor disability over two- to five-year follow-up periods (Bannerman et al 2002, Launer et al 1994, Wannamethee et al 2005). Launer et al (1994) showed that this was especially true in middle-aged, compared to older, women and that although weight change overall resulted in a higher risk of disability than stable weight, those with unintentional weight loss were at higher risk than those with weight gain or intentional weight loss. This could be due to the poorer general health of those with unintentional weight change (Launer et al 1994). Visser et al (1998) considered proportions of body fat in adults aged 65 years and over. They found a dose response association,

with those with higher percentages of body fat being at a higher risk of developing locomotor disability over three years, independently of physical activity levels and chronic illness. There was no evidence regarding the mechanism through which this association could occur.

Two studies in British men found a dose response association between smoking and the onset of locomotor disability, with never smokers being at lowest risk and current smokers at highest risk (Ebrahim et al 2000, Wannamethee et al 2005). Those who stopped smoking longer ago were at lower risk than those who recently gave up. In a Swedish population sample, Ahacic et al (2003) showed that heavy smoking was associated with the onset of locomotor disability. However, Wannamethee et al (2005) also found that four-year onset of locomotor disability was positively associated with giving up smoking during the follow-up period. This may however, be an artefact of the reason for giving up, for example, advice from a doctor due to another condition that may also cause locomotor disability.

Studies of the effect of alcohol consumption on the onset of locomotor disability have found that very heavy (Ebrahim et al 2000, Wannamethee et al 2005) and light levels of drinking (Ebrahim et al 2000, LaCroix et al 1993) can increase the risk of locomotor disability onset. It is unclear exactly what mechanism is creating this pattern, but Wannamethee et al (2005) showed that this association was independent of chronic conditions, suggesting that alcohol is not acting on locomotion through disease.

Low levels of physical activity have been shown to be associated with higher risk of locomotor disability onset (Avlund et al 2002, Ebrahim et al 2000, Wannamethee et al 2005). Wannamethee et al (2005) showed that there was a dose response relationship, with more vigorous activity conferring a larger protective effect. These authors also showed that those becoming more sedentary over the four-year follow-up period were at the highest risk of locomotor disability onset. This may be a direct effect of inactivity. Alternatively, inactivity and the onset of locomotor disability could be the result of an external factor, for example the onset of a chronic condition (Wannamethee et al 2005).

Health conditions and impairments

Many comorbidities have been investigated in relation to the onset of locomotor disability. Studies have found that a simple count of the number of comorbid conditions reported is associated with disability: those with more morbidities are at higher risk of onset (Guralnik et al

1993, Guralnik et al 2001, Wannamethee et al 2005). Other specific diseases and symptoms have also been shown to increase the risk of disability onset.

In the British Regional Heart Study (BRHS), Ebrahim et al (2000) and Wannamethee et al (2005) considered the risks associated with cardiovascular disease (CVD) in men. They found that major CVD (stroke, angina, aortic aneurysm, myocardial infarction) was very strongly associated with the onset of locomotor disability. Similarly, in a group of Americans aged 65 years and over Guralnik et al (1993) found that a history of stroke and high blood pressure increased the risk of locomotor disability onset, as did myocardial infarction or stroke during four-year follow-up. Respiratory diseases and symptoms, including lung disease (Clark et al 1998b, Wannamethee et al 2005), and breathlessness (Guralnik et al 1993, Wannamethee et al 2005) were also associated with an increased risk of onset.

Mental health and cognitive problems have been shown to be associated with the onset of locomotor disability. Clark et al (1998b) showed a 50% increase in the likelihood of locomotor disability over a two-year period in those with poor memory. Guralnik et al (2001) showed a two-fold increase in the catastrophic onset and a two and a half-fold increase in progressive onset of locomotor disability in those with low cognitive function, as defined by the Short Portable Mental Status Questionnaire (Pfeiffer 1975). Avlund et al (2002) showed a higher rate of onset of locomotor disability over a five-year period in 75 year-olds with poor cognitive function. However, Leveille et al (2007) found that psychological symptoms did not mediate the progression of locomotor disability in a three-year follow-up of American women aged 65 years and over.

One of the biggest groups of comorbid illnesses to be investigated in relation to locomotor disability is that of musculoskeletal disease and pain. In the BRHS (Ebrahim et al 2000, Wannamthee et al 2005), arthritis was shown to be strongly related to locomotor disability onset and Guralnik et al (1993) showed a strong association between onset of locomotor disability and hip fracture during follow-up. Associations have also been shown with pain, particularly in the lower limb (Guralnik et al 1993, Wannamethee et al 2005), and when it is bothersome (Clark et al 1998a, 1998b). Leveille et al (2007) showed an approximately two-fold increase in locomotor disability in moderately disabled women with widespread pain. Their findings suggest that this association is direct and the pathway from pain to locomotor disability is not necessarily through reduced activity or reduced muscle strength, as had previously been suggested (for example Dutta & Hadley (1995)).

In addition to these widely investigated factors, diabetes mellitus (Guralnik et al 1993, 2001), onset of cancer during follow-up (Guralnik et al 1993) and visual impairments (Salive et al 1994) have been shown to be related to the onset of locomotor disability. Although not a disease in itself, poor self-rated health has also been associated with the onset of disability (Avlund et al 2002, Guralnik et al 2001). Tiredness in daily activities has been identified as a possible risk factor for the onset of locomotor disability (Avlund et al 2002, 2003), conferring an approximately three-fold increase in the risk of disability over periods of one and a half to five years. Similarly, Fried et al (2000) showed an almost four-fold increase in locomotor disability onset over an 18-month period in women aged 70 to 80 years who reported task modifications at baseline, and an approximate doubling of risk in those who demonstrated slowness in locomotor task performance.

Summary

The factors most commonly and strongly associated with the onset of locomotor disability were older age, low socioeconomic status, high BMI, smoking and low levels of physical activity (Table 3.16). However, chronic disease stood out as the most widely demonstrated factor associated with locomotor disability onset. In particular, Wannamethee et al (2005) showed that the number of chronic diseases was strongly associated with increased reporting of locomotor disability. Particular chronic diseases with a strong association with the onset of locomotor disability were cardiovascular disease, respiratory problems and pain, especially when it is bothersome. This is in keeping with the American attribution study of Leveille et al (2002) in which older women reported that pains in the hip, knee, calf, ankle and foot were the main causes of their lower limb disability. In a similar study in England, Gardener et al (2006) showed that locomotor disability was attributed to pain and breathlessness in both genders.

3.3.5.2 Factors association with recovery from locomotor disability

Many of the factors associated with recovery from locomotor disability are closely related to the factors associated with the onset of disability (Table 3.17). However, it is worth considering these separately, as removal of a factor associated with onset in a person with disability does not necessitate recovery.

Table 3.16 Factors related to the onset of locomotor disability[a]

Category	Factor	Studies providing evidence
Sociodemographic	Female gender	Clark et al (1998a and1998b), Gill et al (2006), Leveille et al (2000),
	Older age	Clark et al (1998b), Guralnik et al (1993, 2001), Gill et al (2006)
	Lower socioeconomic status	Ahacic et al (2000 and 2003), Clark et al (1998a), Ebrahim et al (2000), Guralnik et al (1993), Nordstrom et al (2007) Sainio et al (2007), Wannamethee et al (2005),
Lifestyle	Obesity	Bannerman et al (2002), Clark et al (1998a and 1998b), Ebrahim et al (2000), LaCroix et al (1993), Launer et al (1994), Mendes de Leon et al (2006), Wannamethee et al (2005)
	Tobacco smoking	Ahacic et al (2003), Ebrahim et al (2000), Wannamethee et al (2005)
	Heavy or light alcohol consumption	Ebrahim et al (2000), LaCroix et al (1993), Wannamethee et al (2005)
	Low levels of physical activity	Avlund et al (2002), Ebrahim et al (2000), Wannamethee et al (2005)
Health conditions and impairments	More comorbid conditions	Guralnik et al (1993 and 2001), Wannamethee et al (2005)
	Cardiovascular disease	Ebrahim et al (2000), Guralnik et al (1993), Wannamethee et al (2005)
	Respiratory problems	Clark et al (1998b), Guralnik et al (1993), Wannamethee et al (2005)
	Mental health/cognition problems	Avlund et al (2002), Clark et al (1998b), Guralnik et al (2001)
	Musculoskeletal problems/pain	Clark et al (1998a and 1998b), Ebrahim et al (2000), Guralnik et al (1993), Leveille et al (2007), Wannamethee et al (2005)
	Diabetes Mellitus	Guralnik et al (1993 and 2001)
	Poor self-rated health	Avlund et al (2002), Guralnik et al (2001)
	Tiredness in daily activities	Avlund et al (2001, 2003)

[a]Including only those factors shown by two or more studies to be associated with the onset of locomotor disability

In the BRHS, Wannamethee et al (2005) showed that recovery was associated with a normal BMI (<25kg/m^2). These authors also showed that the likelihood of recovery was reduced by weight gain or weight loss of greater than 10% of original weight over a four-year period. The same was true for weight gain only, in the seven to nine year period prior to baseline.

Physical activity levels were strongly related to locomotor disability in males in the BRHS (Wannamethee et al 2005). This study showed that recovery was more likely in those who were lightly to moderately active at baseline and in those who became active during the follow-up period. Mirroring their findings on onset, these authors found that recovery rates were lowest in those who gave up smoking during the four-year follow-up period.

Very few studies have considered the role of comorbidity in recovery from locomotor disability. Clark et al (1998a) showed that recovery was more likely in those without bothersome pain, lung disease and diabetes mellitus in study of Americans aged 51 to 61 years, whilst Wannamethee et al (2005) showed that recovery was less likely in British men with more chronic conditions.

Table 3.17 Factors related to recovery from locomotor disability

Factor	Studies providing evidence
Female gender	Wannamethee et al (2005)
Older age	Wannamethee et al (2005)
Lower socioeconomic status	Clark et al (1998a)
Obesity	Wannamethee et al (2005)

As with the onset of locomotor disability, the factor most strongly related to recovery from disability was comorbidity. Wannamethee et al (2005) showed that a larger number of comorbid diseases were associated with a lower odds of recovery in British males. However these authors failed to show associations with several individual chronic diseases, for example diabetes, arthritis and cancer.

3.4 Summary and discussion

3.4.1 Principal findings

The picture of locomotor disability in middle- and old-age that this review paints is somewhat disjointed due to methodological differences between studies. Nevertheless, what emerges relatively clearly is that at any point in time, however it is defined and in any geographical location, locomotor disability may be reported by a substantial proportion of adults aged 50 years and over

in the general population (2% to 65%). The prevalence of locomotor disability appears to increase with age and to be higher in females.

One study, using repeated assessment of locomotor disability with short intervals between these assessments, has shown that it may be quite a dynamic state that people can and do move in and out of in the course of five years (Gill et al 2006). Such a view emphasises not just the importance of determinants of acquiring locomotor disability but also of recovering from it. This finding gives greater weight to the argument of Sakari-Rantala et al (2002) that the use of a dichotomy to define the presence of locomotor disability may artificially suggest large differences in disability levels over time because small fluctuations in disability, be these real or due to measurement error, in those whose ability lies close to the threshold between function and disability, result in their changing categories.

Locomotor disability can be assessed by self-report or performance-based indicators. Whilst many performance-based assessments provide measurement, as define in this thesis, there has previously been no such measure derived from self-reported data. As discussed in Section 3.3.1.2, several studies have directly compared observed functional performance and self-reported measures and found variable levels of comparability. A lack of comparability might result from the incorporation of health expectations and a societal perspective into the self-report of disability. However, the two types of measure usually cover different task difficulties or parameters, with self-report measures generally able to assess higher levels of functioning than performance measures. For example, Sainio et al (2007) compared observed time taken to complete climbing and descending two steps with self-perceived limitation in climbing a flight of stairs, whilst Bohannon et al (2004) compared gait speed over 7.62 metres with self-reported ability to walk more than a mile. Hence, in using performance measures to assess more difficult tasks, one must extrapolate beyond what is actually observed. There is evidence though that such performance-based measures can identify disability that is not yet perceived as such by the respondent (Guralnik et al 2000, Onder et al 2005, Brach et al 2007, Huang et al 2010). In making a choice between performance-based and self-report measures, these issues must be considered, as must the costs involved: the cost involved in assessing performance will usually be higher than would be the case for self-report. Furthermore, there are practical issues in the collection of performance data in large samples. Whilst some studies, such as the Established Populations for Epidemiologic Studies of the Elderly (Melzer et al 2003) have measured performance in large numbers of people, in other

studies, this may not be possible. Given the considerations of the practicality of the use of performance measures, and the lack of a true measure of locomotor disability from self-report instruments, this thesis will concentrate on self-report instruments.

3.4.2 Strengths and limitations of this review

The search strategy for this review was clearly defined and based on previously used criteria (Stuck et al 1999). The reference lists of these papers were then hand searched for additional studies that may have been missed by the electronic search due to the plethora of terms used to describe locomotor disability. Despite this conservative approach to the searching of the literature, it is possible that some studies have been missed.

An additional limitation of this review was the lack of comparability between studies. A large number of different definitions of locomotor disability have been used, making comparison across studies problematic. There is little consensus regarding which activities come under the umbrella of locomotion, and in which or how many of these activities a person must be disabled in order to have 'locomotor disability'. Even if studies used the same locomotor activities to define locomotion, questions were rarely worded or responses scored in the same way. Without a standard tool with which to measure or even assess locomotor disability, it is not possible to make cross-study comparisons. This issue of definition is particularly important given the evidence surrounding response category cut point shifts (Melzer et al 2004) and other non-physiological reasons that self-reported levels of locomotor disability differ.

Many factors have been shown to be associated with the onset of locomotor disability: older age, female gender and lower socioeconomic status. Similar evidence exists for obesity, inactive life-styles and heavy smoking and drinking. Several groups of comorbid diseases including cardiovascular, mental health and musculoskeletal conditions have also shown strong associations with onset. However, only normal BMI, high levels of physical activity and a lack of comorbid disease have been associated with increased levels of recovery.

An association in a longitudinal study does not guarantee causality or remove the possibly of reverse causality, i.e. the disability causing the other factor. It is also possible that some of the factors associated with locomotor disability in the general population are new to the field and have only been investigated in cross-sectional studies: these factors will be missed using this approach. Some studies have considered respondent attribution, i.e. what the respondent thinks caused or is

causing their locomotor disability. This is different to studies of association in that the individual's view on the cause of their disability is described, as opposed to their being asked to respond to questions regarding specified factors. This may be important, particularly if the individual attributes the cause of their disability to something that would not normally be assessed in an association study or in a visit to a clinician. It is also possible however, that attribution studies, and perhaps to a lesser extent association studies, may not find factors that are directly associated with locomotor disability as they operate through some other mechanism. Despite the obvious differences between association and attribution studies, Gardener et al (2006) drew similar conclusions about the effects of lower limb pain and dyspnoea on locomotor disability in their attribution study as many of the association studies. It is therefore unlikely that the strategy of considering only longitudinal studies of association has missed a large number of factors to which people attribute their locomotor disability.

3.4.3 Implications of the findings of this review

This review has identified two major weaknesses in the current state of the measurement of locomotor disability: the lack of an agreed upon definition of what tasks or activities constitute locomotor disability and the lack of true measurement (as defined in Chapter 2) with a self-report instrument, to allow a full investigation of the epidemiology of locomotor disability.

Feinstein et al (1986) and de Vet et al (2003) have criticised the plethora of instruments available to assess health status. Therefore, it would be preferable for an agreed upon measure of locomotor disability to come from items of a pre-existing instrument, rather than further new items being created to assess this concept. The SF-36 Physical Functioning subscale (PF-10) is a widely-used, self-report measure that has been cited in this context and has been used in some studies to assess locomotor disability (Bohannon et al 2004, Peat et al 2006a). It contains walking and stair-climbing items similar to those used in many of the studies included in this review. Whilst the studies in this review have used a range of items, most have been derived specifically for the study or have used items from a range of other questionnaires. Hence the PF-10 has unqiue potential, in terms of its widespread use and previous psychometric evaluations, and contains relevant items with which to consider the derivation of a new measure of locomotor disability. Furthermore, the PF-10 has recently been asserted by Syddall et al (2009) to represent a valid

measure of locomotor disability in epidemiological studies, although the evidence for this claim was limited.

Chapter 6 will consider the PF-10 as a potential measure of locomotor disability. First, Chapter 4 describes the North Staffordshire Osteoarthritis Project (NorStOP), which provides the majority of the data used in this thesis, and evaluates its suitability for the purpose. Chapter 5 then considers the necessary properties for a good measurement instrument.

4 The North Staffordshire Osteoarthritis Project (NorStOP)

4.1 Introduction

In order to fulfil the objectives set out in Chapter 1, a population-based dataset was needed in which suitable data, including the PF-10, had been collected. A source of such data was the North Staffordshire Osteoarthritis Project (NorStOP). This was a population-based prospective cohort study conducted via postal survey. The main aim of the NorStOP was to study the clinical syndrome of osteoarthritis in a general population sample of older people (Thomas et al 2004a). Participants also provided information on a range of health and health-related concepts.

This chapter will discuss the various stages of the study and its contents (Section 4.2), as well as response to these various stages (Section 4.3) and the cohort's consequent representativeness of the general population (Section 4.4). Finally, Section 4.5 considers the overall suitability of the NorStOP data to fulfil the objectives of this thesis, highlighting potential strengths and weakness of the dataset in this respect.

4.2 Study design

The target population for the NorStOP was community-dwelling adults aged 50 years and over. These people were selected via general practice lists in North Staffordshire and recruited over three time-periods, giving three subcohorts (NorStOP1, NorStOP2 and NorStOP3). A two-stage mailing strategy ("Health Survey" questionnaire and "Regional Pains Survey" questionnaire) was employed. This section describes details of the study design, mailing process and content of the questionnaires, as relevant to this PhD project. The majority of this information has been published previously (Thomas et al 2004a).

4.2.1 Pilot study

A pilot study was completed during October and November 2001. This consisted of a random sample of 500 adults aged 50 years and over from a single general practice in North

Staffordshire. The practice is a member of the Keele General Practice Research Partnership (KGPRP). The sample was checked by the general practitioner (GP) for exclusions (for example severe psychiatric or terminal illness).

Members of the sample were sent a Health Survey questionnaire and, if they consented to further contact from the researchers following this first questionnaire, a repeat Health Survey questionnaire was also sent. This allowed the repeatability of the questionnaire to be assessed.

Each questionnaire that was mailed with a letter from the general practice, a study information leaflet and a prepaid and addressed return envelope. After two weeks, those who had not responded to the questionnaire received a postcard reminder. Those who had not returned the questionnaire after a further two weeks received a second copy of the questionnaire and reminder letter. This reminder process has been used in previous studies from the Research Centre (Jinks et al 2001, Boardman et al 2003) where it has been found to be acceptable and to contribute substantially to response rates.

Throughout the mailing process, checks for deaths and departures from the general practices were carried out by KGPRP staff in order to maintain up-to-date records regarding participants.

4.2.2 Recruitment into the NorStOP subcohorts

The first subcohort, NorStOP1, was recruited during March and April 2002. All adults aged 50 years and over at three general practices in North Staffordshire were identified by KGPRP informatics staff. The GPs then screened the list of potential participants for exclusions (as per the pilot study).

Those who responded to this questionnaire, gave consent to be contacted again and indicated that they had experienced hand pain or problems, hip pain, knee pain or foot pain in the last year, were sent a Regional Pains Survey questionnaire. For both the Health Survey and the Regional Pains Survey, those people who did not respond to the questionnaire after two weeks were sent a reminder postcard. As in the pilot study, if no response was received after a further two weeks, a second copy of the questionnaire was sent.

The date of birth and gender given on the returned questionnaires were checked against data from the general practice list to ascertain that the correct person had completed each

questionnaire. If the details did not match, names and addresses were used to decide whether the data could be assumed to be from the intended recipient.

As in the pilot study, checks were carried out throughout the mailing process by KGPRP staff for deaths and departures from the practice and a record was kept of those who, based on their response to the Health Survey questionnaire, consented to be contacted again for future follow-up of the cohort.

The same procedures were followed to recruit the NorStOP2 subcohort from three further general practices between July 2002 to June 2003 and the NorStOP3 was recruited from two general practices from February 2004 to February 2005. In addition to recruiting participants to a large, survey-based cohort, an objective of the NorStOP2 was to provide a sampling-base for a clinical epidemiology study of knee pain (CAS-K; Peat et al 2004), whilst NorStOP3 provided a sampling-base for a parallel clinical epidemiology study of hand pain and problems (CAS-HA; Myers et al 2007). To recruit these two subcohorts, those people who responded to the Regional Pains Survey questionnaire and reported knee (hand) pain were invited to attend the local hospital for a clinical assessment with a research therapist. Those people who attended underwent a clinical interview, assessment, and plain radiographs of the knees and hands. In the CAS-HA, this assessment included the Short Physical Performance Battery (SPPB) (Guralnik et al 1994), which is used in Chapters 6 and 12 and is described in full in Appendix B.

In this thesis, NorStOP2 will refer to participants who responded to the postal surveys, regardless of whether they subsequently became part of the CAS-K, and NorStOP3 will refer to participants who responded to the postal surveys, regardless of whether they subsequently became part of the CAS-HA.

4.2.3 Follow-up of the NorStOP1 subcohort

4.2.3.1 Three-year follow-up

Those people who had consented to further contact following the recruitment phase of the NorStOP1 were eligible for follow-up three years later. Health informatics staff from KGPRP checked the current practice registers to identify all those in this sample that were no longer registered with the practice prior to the mailing of the three-year follow-up questionnaires. For those no longer registered with their original general practice, the National Health Service Strategic Tracing Service (NHSSTS) was used to ascertain either the current contact details for the person

and their new GP, or that the person had died. These checks for deaths and departures continued at two-weekly intervals throughout the mailing period. Those identified as having left the practice during this mailing period were also followed-up using the NHSSTS. In a small number of cases, the NHSSTS returned the same GP and address information as in the study database, which was incorrect. These people could no longer be followed-up as part of the NorStOP.

The same two-stage mailing procedure used at recruitment was used at three-year follow-up. All cohort members were sent a Health Survey questionnaire along with a letter from their general practice and a patient information sheet. Respondents to this questionnaire who consented to further contact and reported hand pain or problems, hip pain, knee pain or foot pain, were sent a Regional Pains Survey questionnaire. Reminders for each stage of the survey were sent to non-responders after two and four weeks.

4.2.3.2 Six-year follow-up

Those people who had consented to further contact following the three-year follow-up phase of the NorStOP1 were eligible to be followed-up again after a further three years (six-year follow-up). Similarly to the three-year follow-up, prior to the mailing of the six-year follow-up questionnaires, KGPRP staff identified those people no longer registered with the participating practices and these people were traced via the NHSSTS. However, unlike during the three-year follow-up, patients found to be no longer registered with the practice during the mailing process were not traced using the NHSSTS. This decision was made as its application in the three-year follow-up was found to excessively lengthen the mailing process and did not substantially increase the number of survey responses.

The same mailing procedure was used at six-year follow-up as at recruitment and three-year follow-up. All cohort members were sent a Health Survey questionnaire along with a letter from their general practice and a patient information sheet. Respondents to the Health Survey who consented to further contact and reported hand pain or problems, hip pain, knee pain or foot pain were sent a Regional Pains Survey questionnaire. Reminders were sent to non-responders after two and four weeks.

4.2.4 Follow-up of the NorStOP2 and 3 subcohorts

The NorStOP 2 and 3 subcohorts were also followed-up three and six years after their inception. However, these data were not available within the time frame of this thesis and so these follow-ups will not be discussed further.

4.2.5 Health Survey questionnaire content

The Health Survey questionnaire collected information on a variety of health and health-related concepts. These included general health; participation; social networks; mental health; perceptions of illness; occupational, demographic and lifestyle factors; morbidity; bodily pain and physical functioning. During the follow-up stages of the NorStOP, some questionnaire content was changed. A copy of the baseline Health Survey questionnaire is included in Appendix B. Three- and six-year follow-up questionnaires were similar in terms of the data used in this thesis. Of specific interest in this thesis are basic demographic characteristics, individual socioeconomic status, pain status and physical function. Details of how each of these concepts was assessed and used in this thesis are given below, along with details of other data from the NorStOP used in this thesis.

4.2.5.1 Basic demographic characteristics

Age was calculated from the date of birth given in the baseline Health Survey questionnaire (Appendix B, page 19, question 1). This date may have been different from that on the database received from the general practice, but once it was established that the correct person had completed the questionnaire (see Section 4.2.2), the information given by the individual was taken as the correct date of birth. Where possible, the name of the respondent from the GP list was used to ascertain gender if the data from the GP list and the baseline Health Survey questionnaire (Appendix B, page 20, question 2) were contradictory.

Living arrangement was assessed at baseline using the question, "*Do you live alone?*". Possible responses were "*Yes*" or "*No*" (Baseline Health Survey questionnaire (Appendix B): page 20, question 4).

Ethnic origin was self-reported in the baseline Health Survey questionnaire (Appendix B: page 22, question 16), as "*White UK/European*", "*Afro Caribbean*", "*Chinese*", "*Asian*", "*African*" or "*Other*".

4.2.5.2 Socioeconomic status

Individual socioeconomic status was measured in three domains traditionally used in older adults: education, occupational class and perceived adequacy of income (Grundy & Holt 2001).

Educational attainment was assessed at baseline using the question "*Did you go on from school to full time education or university?*". Possible responses were "*Yes*" ('further education') and "*No*" ('school-age education only') (Baseline Health Survey questionnaire (Appendix B): page 22, question 13).

Occupational class was assessed at baseline using current job title (most recent job for those who were not working) (Baseline Health Survey questionnaire (Appendix B): page 20, question 6). Job titles were classified according to the Standard Occupational Classification (Office for National Statistics 2000, Office for National Statistics 2002). These were then regrouped into 'manual' (lower supervisory/technical, semi-routine occupations, routine occupations) and 'non-manual' (higher managerial, higher professional, lower managerial/professional, intermediate occupations) occupations for use in analysis. Self-employed people were excluded from this categorisation, as it was not clear into which group they should be placed.

Perceived adequacy of income was also assessed at baseline, using the item "*Thinking about the cost of living as it affects you, which of these descriptions best describes your situation?: Find it a strain to get by from week to week; Have to be careful with money; Able to manage without much difficulty; Quite comfortably off*" (Thomas 1999) (Baseline Health Survey questionnaire (Appendix B): page 22, question 15). For the purposes of analysis, these responses were dichotomised into 'inadequate' (find it a strain to get by from week to week, have to be careful with money) and 'adequate' (able to manage without much difficulty, quite comfortably off).

4.2.5.3 Pain

Pain was assessed at baseline using a single item and a body manikin. The item asked, "*In the past 4 weeks, have you had pain that has lasted for one day or longer in any part of your*

body?" with possible responses of "*Yes*" and "*No*" (Baseline Health Survey questionnaire (Appendix B): page 27, question 1). Respondents were asked to include "*any ache, discomfort or stiffness*" in their concept of pain and were asked to exclude pain caused by feverish illness, such as flu, and menstrual pain. This assessment of pain was repeated in the Health Survey questionnaire at three- and six-year follow-ups.

Front and back views of the body manikin were displayed and respondents reporting pain in the single item were asked to shade where they had pain. Standard transparent templates with the borders marked were used to assess in which area of the body respondents had reported pain. This method has been shown to be repeatable (Lacey et al 2005). The templates split the manikins into 44 mutually exclusive areas (Figure 4.1a) with additional site-specific definitions of the hip (Birrell et al 2000) (Figure 4.1b) and low back (Papageorgiou et al 1995) (Figure 4.1c).

The following rules were applied to combine the screening item and body manikin for use in Chapter 8:

- Those responding negatively to the pain item and not shading the manikin were assigned to a 'no pain' group.
- Those responding positively to the pain item and shading the manikin were considered to have pain in the shaded areas.
- Those people who responded "*Yes*" to the pain item but did not shade any areas on the manikin were excluded from analyses.
- Those who responded "*No*" to the pain item but shaded at least one area on the manikin were excluded from analyses.
- Those people who did not complete the screening item were excluded from analyses.

This created two unambiguous groups for use in the ordinal regression analyses presented in Chapter 8.

Due to attrition from the cohort during the six years of follow-up and the timeframe of this thesis only allowing follow-up data from the NorStOP1 to be used, the use of the manikin and related screening item were reconsidered for use in Chapter 14. The aim being to keep as 'clean' a group as possible, whilst maximising the numbers of people available for the analysis. It was decided to apply the following rules to combine the screening item and the body manikin.

- Respondents who shaded the manikin were assumed to have pain in the shaded areas, irrespective of response to the screening item.

- Respondents who did not shade the body manikin and responded "*No*" to the screening item were assumed not to have pain.
- Respondents who did not shade the body manikin and responded "*Yes*" to the screening item were excluded from analyses.
- Respondents who left the screening item blank were excluded from analyses, irrespective of manikin shading.

Full details of this consideration process are given in the Appendix B.

4.2.5.4 Physical function

Physical function was assessed using the Physical Functioning subscale (PF-10) of the SF-36 Health Survey Questionnaire (Ware & Sherbourne 1992), which was included in the NorStOP at all time points (for example, Baseline Health Survey questionnaire (Appendix B): page 30, questions a to j). A detailed review of the measurement properties of the PF-10 is presented in Chapter 6.

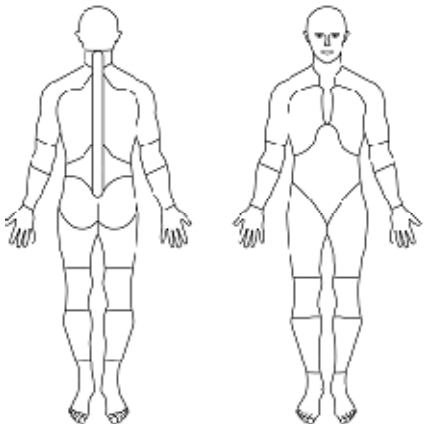
4.2.5.5 Other health-related concepts

Participants were asked to give their ***height*** and ***weight*** using either metric or imperial units at each time point (for example, Baseline Health Survey questionnaire (Appendix B): page 21, questions 1 and 2). Values of height and weight were used to calculate ***body mass index*** in kilograms per metre^2 (kg/m^2). Participants were also asked to state whether they had certain ***health conditions*** (for example, Baseline Health Survey questionnaire (Appendix B): page 23, questions 1 and 2) and to rate their ***general health*** (for example, Baseline Health Survey questionnaire (Appendix B): page 3, question 1).

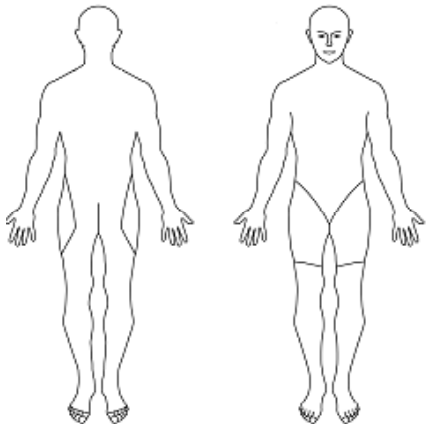
Participation restriction was assessed using the Keele Assessment of Participation (KAP) (Wilkie et al 2005) (for example. Baseline Health Survey questionnaire (Appendix B): pages 5 to 7, questions 1 to 11). This instrument allows the assessment of a responder in terms of their participation restriction in 11 separate areas of life, and also provides a total number of areas of restriction.

Figure 4.1 Areas of the body manikin

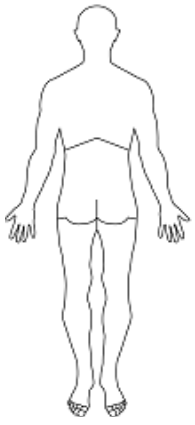
a. 44 mutually exclusive areas



b. The hip (Birrell et al 2000)



c. The low back (Papageorgiou et al 1995)



Anxiety and ***depression*** were assessed using the Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith 1983) (for example, Baseline Health Survey questionnaire (Appendix B) 13 to 15, questions 1 to 14). ***Cognitive complaint*** was assessed using the Sickness Impact Profile (SIP) Alertness subscale (Bergner et al 1981) (for example, Baseline Health Survey questionnaire (Appendix B): page 24, question 3). The KAP, HADS and SIP Alertness scales were included in all stages of the NorStOP and responses to these scales are used in Chapters 6 and 12.

At baseline only, participants were asked about the ***frequency with which they undertook certain activities*** (Baseline Health Survey questionnaire (Appendix B): pages 8 and 9, questions a to u). Responses to these items were dichotomised for use in Chapters 6 and 12: 'often' ("*all days*", "*most days*") and 'less often' ("*some days*", "*few days*", "*no days*"). Again at baseline only, respondents were asked whether they used ***aids or needed assistance from others*** to move around inside and outside their homes (Baseline Health Survey questionnaire (Appendix B): page 10, questions 2 and 3). Possible responses to these items were "*Yes*" and "*No*".

4.2.6 Ethical approval and informed consent

Ethical approval was gained separately for the baseline stage of the NorStOP1 (including the pilot study), and for the NorStOP2 and 3, which were gained together from the North Staffordshire Local Research Ethics Committee (REC). REC numbers 1351 and 1430 respectively. Ethical approval for the recruitment stage of CAS-K and the CAS-HA was included in the approval for the NorStOP2 and 3. For the NorStOP1 three-year follow-up, ethical approval was gained separately (REC number 05/Q2604/20) and a substantial amendment was made to this application for the six-year follow-up.

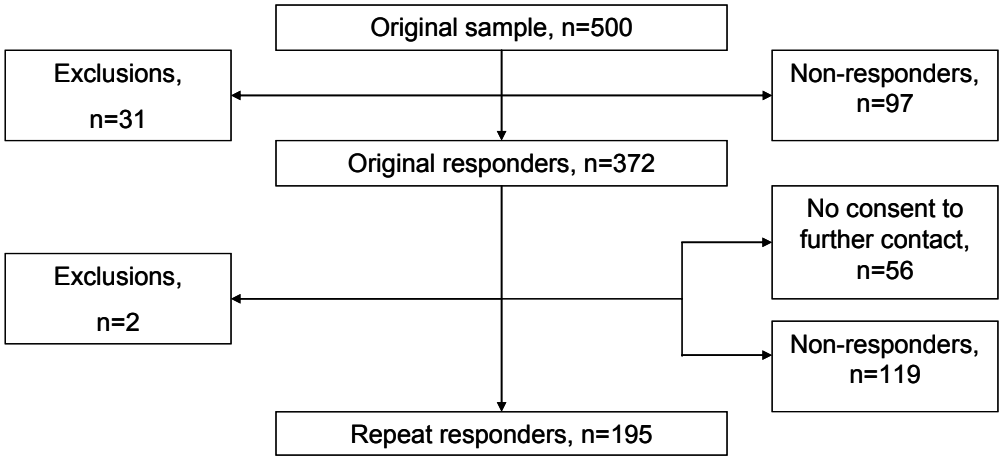
The Health Survey questionnaires used for each stage included a page that collected signed consent from the individual (for example, Baseline Health Survey questionnaire (Appendix B): page 33). At baseline, participants completing the Health Survey questionnaire were asked to complete a consent form giving permission (or not) for researchers to contact them again in the future and for researchers to access their medical records. At three- and six-year follow-up, the consent form gained consent for further contact only.

4.3 Response to the surveys

4.3.1 Pilot study

Of the 500 people randomly selected to take part in the pilot study of the Health Survey questionnaire, 372 (adjusted response[1] 79%) responded to the original questionnaire and 316 of these were mailed a repeat Health Survey questionnaire. Completed questionnaires were received from 195 of these people (adjusted response 62%) (Figure 4.2).

Figure 4.2 Flow diagram of participants in the NorStOP pilot study



4.3.2 The NorStOP subcohorts

KGPRP staff identified 26,705 (NorStOP1: 11,309; NorStOP2: 8,984; NorStOP3: 6,412) members of the eight general practices aged 50 years and over, of whom 26,129 were eligible to take part in the NorStOP and were mailed the Health Survey questionnaire. Of these people, 18,497 (NorStOP1: 7,878 (Thomas et al 2004b); NorStOP2: 6,108 (Peat et al 2006b); NorStOP3: 4,511 (Marshall et al 2009)) (adjusted response rate 71%) responded to the Health Survey questionnaire. After this baseline stage of the study, 12,641 (NorStOP1: 5,366 (Thomas et al 2007); NorStOP2: 4,091; NorStOP3: 3,184) people were eligible for the three-year follow-up.

---

[1] Adjusted response refers to the response rate calculated with those who were excluded from the mailing process, for example through death, stroke, dementia or incorrect address information, removed from the denominator

4.3.3 Follow-up of the NorStOP1 subcohort

4.3.3.1 Three-year follow-up

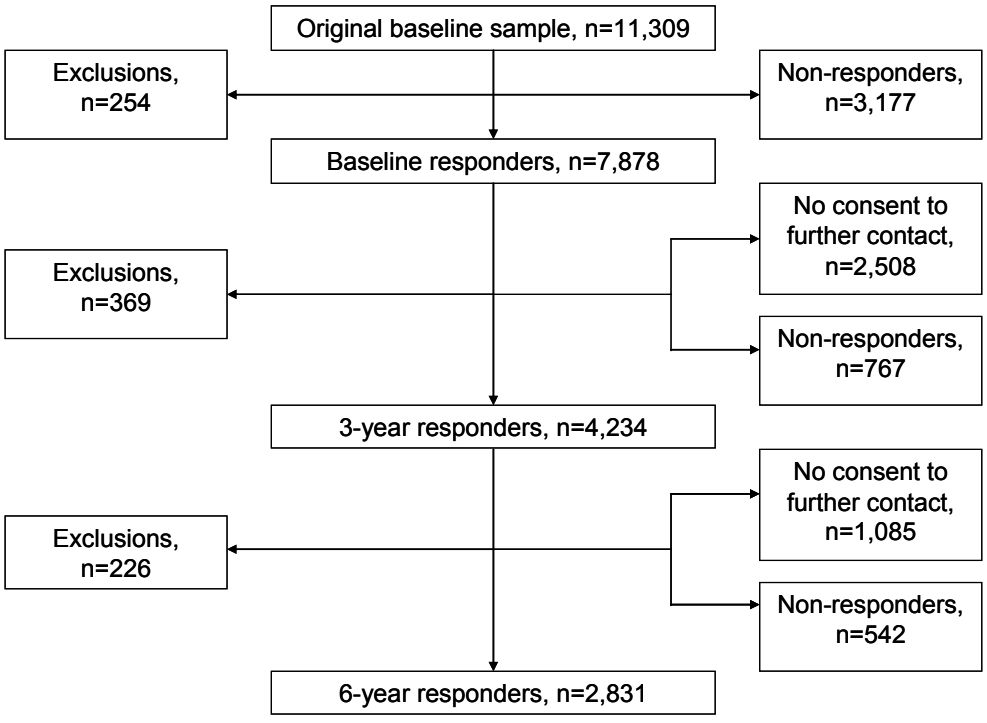
Immediately before the three-year follow-up, 5,001 people in the NorStOP1 were identified as still being eligible for follow-up. Completed Health Survey questionnaires were received from 4,234 people (adjusted response rate 85%) (Thomas et al 2007). At the end of the three-year follow-up stage, 3,596 people in this subcohort were eligible for the six-year follow-up.

4.3.3.2 Six-year follow-up

Immediately before the six-year follow-up, 3,373 people were identified as still being eligible for follow-up. Completed Health Survey questionnaires were received from 2,831 people (adjusted response rate 84%).

Figure 4.3 shows the overall response over time to the Health Survey questionnaires in the NorStOP1.

Figure 4.3 Response to Health Survey questionnaires in the NorStOP1 subcohort over six years



4.3.4 Completeness of the data

Table 4.1 shows the levels of missing data associated with key variables from the NorStOP that will be used in this thesis. There were some instances of high levels of missing data within certain constructs used in the analyses presented in this thesis. As might be expected, levels of missing data were higher where multiple items were required in order to construct a score for a construct, for example the SF-12 physical and mental health component scores.

4.4 Representativeness of the NorStOP cohort

4.4.1 Responders versus non-responders at baseline

Basic demographic information relating to those people who did not respond to the survey at baseline was available from the general practices. Table 4.2 shows the age distribution, by gender, of the Staffordshire population and the total mailed NorStOP sample at recruitment, broken down into responders, exclusions and non-responders.

Those aged 80 years and over made up a much larger proportion of the mailed sample than they did the Staffordshire population in both genders. There was a correspondingly lower proportion of people in the youngest age-group in the mailed sample than in the Staffordshire population.

The age distribution of female responders is generally similar to that of the mailed sample, although there is some under-representation of those aged 80 years and over, as the level of exclusions was higher at older ages. The youngest age-group is somewhat under-represented in male responders. More males in this age-group were excluded than at older ages. Non-response was higher at younger ages in both genders.

Overall, there is some evidence that the mailed sample for the NorStOP may not be wholly representative of the Staffordshire population in terms of age and gender. This is not necessarily surprising, given that the NorStOP was sampled exclusively from the North of the county. Overall, responders to the NorStOP at baseline were generally older than the Staffordshire population.

Table 4.1 Levels of missing data among items and scales within the NorStOP at baseline, and at three- and six-year follow-up in the NorStOP1

%	NorStOP Baseline (n=18,497)	NorStOP1 three-year follow-up (n=4,234)	NorStOP1 six-year follow-up (n=2,831)
Educational attainment	2.6	N/A	N/A
Occupational class[a]	15.0	N/A	N/A
Perceived adequacy of income	2.7	N/A	N/A
SF-12 PCS	14.0	12.1	10.8
SF-12 MCS	14.0	12.1	10.8
PF-10 scale score	1.7	2.0	1.8
Pain manikin (Chapter 6 usage)	11.6	11.8	3.1
Pain manikin (Chapter 8 usage)	6.0	17.6	4.0

N/A - data not used in this thesis; [a]percentage missing includes those excluded from classification as self-employed

4.4.2 The NorStOP compared to the local and national population

In order for the results of analyses carried out in the NorStOP cohort to be generalisable to a wider population, participants in the NorStOP should be representative of such a population. This section will compare the NorStOP cohort, i.e. those who responded to the Health Survey questionnaire at baseline, with the English, and where possible, the Staffordshire population on a number of key demographic characteristics and those attributes of particular interest in this thesis.

Table 4.2 Age and gender distribution of the Staffordshire population, the mailed NorStOP sample, the responding NorStOP cohort, exclusions and non-responders at baseline

	Staffordshire (%)		Mailed sample (%)		Responders (%)		Exclusions[a] (%)		Non-responders[b] (%)	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
50 to 59 years	45.2	38.1	38.4	32.2	33.5	31.0	42.8	19.4	48.8	36.2
60 to 69 years	31.2	27.6	31.0	27.0	32.8	29.1	24.1	14.8	27.5	22.3
70 to 79 years	18.6	22.4	21.3	23.7	24.2	25.6	16.9	19.0	15.4	19.0
80 years and over	4.9	11.9	9.3	17.0	9.4	14.3	19.2	46.8	8.4	22.4

[a]Exclusions included those who died, were identified as having dementia, suffered a stroke or not living at the address to which the questionnaire was mailed. These people were therefore unable to complete the questionnaire; [b]Non-responders included those who did not return a questionnaire, returned a blank questionnaire and those who contacted the Research Centre to remove themselves from the study, but were not eligible to be excluded

4.4.2.1 Age

Table 4.3 shows the age structure of the English and Staffordshire populations, the NorStOP mailed sample and the responding cohort. The age structure of England is mirrored by the age structure of the Staffordshire population. However, as described in Section 4.4.1, the mailed NorStOP sample was slightly older than this Staffordshire population, with more people aged over 65 years and fewer people below the age of 55 years. There is some evidence of response bias by age, with those aged 60 to 79 years more likely to respond than the younger or older participants, but the majority of the difference in age structure between the NorStOP and England is likely to be due to the sampling frame, rather than a bias in the response to the survey.

Table 4.3 Age structure of the English and Staffordshire populations, the NorStOP mailed sample and the NorStOP responding cohort

	%			
	England[a]	Staffordshire[a]	Mailed NorStOP sample	NorStOP responders[b]
50 to 54 years	21.1	21.9	15.7	13.9
55 to 59 years	17.3	18.7	19.4	18.3
60 to 64 years	14.9	15.4	15.0	15.5
65 to 69 years	13.4	13.3	13.8	15.3
70 to 74 years	12.0	11.8	12.2	13.6
75 to 79 years	10.0	9.4	10.4	11.3
80 to 84 years	6.5	5.8	7.7	7.7
85 to 89 years	3.4	2.8	3.8	3.3
90 years and over	1.4	1.1	2.0	1.2

[a]Calculated from 2001 Census: standard tables downloaded from Nomis https://www.nomisweb.co.uk/home/Census2001.asp, 7 March 2008; [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire

4.4.2.2 Gender

Table 4.4 shows the gender structure of the English and Staffordshire populations, the NorStOP mailed sample and the responding cohort by age-group. The gender structure of England is mirrored by the gender structure of the population in Staffordshire. In the mailed sample for the NorStOP, there was a higher proportion of females at older ages than in Staffordshire or England. In the cohort of responders to the NorStOP, the proportion of females was higher than in the mailed sample in all age-groups up to age 75 years.

Table 4.4 Gender distributions of the English and Staffordshire populations, the NorStOP mailed sample and the NorStOP responding cohort, by age-group

	% female			
	England[a]	Staffordshire[a]	Mailed NorStOP sample	NorStOP responders[b]
50 to 54 years	50.5	49.6	49.8	54.8
55 to 59 years	50.5	49.9	49.7	53.1
60 to 64 years	50.9	50.1	49.9	52.3
65 to 69 years	52.0	51.9	51.7	53.2
70 to 74 years	54.5	54.4	56.3	57.1
75 to 79 years	58.0	57.6	57.3	57.2
80 to 84 years	62.2	62.8	64.7	62.8
85 to 89 years	67.9	68.5	69.8	67.6
90 years and over	73.8	74.0	80.2	78.2

[a]Calculated from 2001 Census: standard tables downloaded from Nomis https://www.nomisweb.co.uk/home/Census2001.asp, 7 March 2008; [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire

4.4.2.3 Ethnicity

A higher proportion of the Staffordshire population and NorStOP responding cohort is white than in England as a whole in each age-group (Table 4.5). Although the NorStOP cohort may not be representative of England, it is broadly representative of the population in Staffordshire, thus reducing the likelihood of response bias due to ethnicity. The effect of any such a bias is likely to be negligible due to the small proportion of non-white people overall in these age-groups.

4.4.2.4 Living arrangement

In both genders, the proportion of people living alone is lower in Staffordshire than in England as a whole and lower still in the NorStOP responding cohort (Table 4.6). The difference between the population data and the NorStOP may be as a result of the different items used to ascertain this information in the NorStOP and in the 2001 Census, from which the population data are taken. The Census data are based on those who are living in a couple, whereas the NorStOP data refer to living with any other person, be this a spouse/partner (as in the Census data) or another person, for example a sibling or child. For this reason, it is not possible to quantify the extent of any response bias related to living arrangement.

Table 4.5 Ethnicity distribution of the English and Staffordshire populations and the NorStOP responding cohort, by age-group

	% white		
	England[a]	Staffordshire[a]	NorStOP[b]
50 to 54 years	94.9	98.6	99.3
55 to 59 years	95.6	98.9	99.3
60 to 64 years	94.7	98.7	99.4
65 to 69 years	95.5	98.8	99.5
70 to 74 years	96.7	99.9	99.6
75 to 79 years	97.9	99.2	99.4
80 to 84 years	98.2	99.2	99.6
85 to 89 years	98.8	99.5	99.1
90 years and over	98.6	99.5	100.0

[a]Calculated from 2001 Census: standard tables downloaded from Nomis https://www.nomisweb.co.uk/home/Census2001.asp, 7 March 2008; [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire

Table 4.6 Distribution of living arrangement in the English and Staffordshire populations and the NorStOP responding cohort, by gender and age-group

	% living alone		
	England[a]	Staffordshire[a]	NorStOP responders[b]
Males			
50 to 59	20.8	17.4	12.6
60 to 69 years	20.7	17.3	13.5
70 years and over	31.8	30.7	18.2
Females			
50 to 59	24.3	18.7	13.7
60 to 69 years	33.4	29.5	23.3
70 years and over	65.2	64.0	38.8

[a]Calculated from 2001 Census: standard tables downloaded from Nomis https://www.nomisweb.co.uk/home/Census2001.asp, 7 March 2008). Living alone defined as not with spouse or partner; [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire. Living alone defined as not living with any other person

4.4.2.5 Educational attainment

In both genders and all age-groups, the proportion of people who have attended further education is lower in Staffordshire than in England as a whole (Table 4.7). The proportion of people

who have attended further education is substantially lower in the NorStOP than in Staffordshire. This is particularly true in males.

This difference in educational attainment between Staffordshire and the NorStOP is likely to be due to the different definitions used to assess educational attainment in the NorStOP and the Census. The NorStOP Health Survey questionnaire asked participants to state whether they had attended full time education or university after leaving school whereas the Census data was concerned with the highest level of qualification gained. Hence school age education only in the NorStOP was equated in the Census data to having no qualifications, only one GCSE-level qualification, an NVQ level 1 or a foundation GNVQ. Gaining a higher level of qualification, for example more than one GCSE-level qualification, does not automatically equate to having attended further education. Hence this comparison is rather coarse, and it is difficult to ascertain the level of any response bias in terms of educational attainment. Due to the difference in definitions, it seems likely that the real difference in educational attainment is smaller than suggested by Table 4.7.

Table 4.7 Distribution of further education in the English and Staffordshire populations and the NorStOP responding cohort, by gender and age-group

	% with further education		
	England[a]	Staffordshire[a]	NorStOP responders[b]
Males			
50 to 59 years	52.4	47.6	15.4
60 to 69 years	41.0	35.6	8.1
70 years and over	35.8	29.0	5.7
Females			
50 to 59 years	44.5	39.2	20.0
60 to 69 years	31.7	25.5	12.9
70 years and over	27.5	20.5	9.5

[a]Calculated from 2001 Census: standard tables downloaded from Nomis https://www.nomisweb.co.uk/home/Census2001.asp, 7 March 2008. Further education defined as ONS classifications level 2 or above (5+ O levels, 5+ CSEs (grade1), 5+ GCSEs (grade A*-C), School certificate, 1+ AS levels, NVQ level 2, Intermediate GNVQ or equivalents). School education only defined as ONS classification 'no qualifications' or level 1(1+ O levels/CSEs/GCSEs (any grade), NVQ level 1, Foundation GNVQ); [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire. Further education defined as having gone on to full time education of university after leaving school.

4.4.2.6 Employment status

The proportion of people at older ages in employment in Staffordshire is slightly lower than in England as a whole (Table 4.8), although broadly similar proportions were retired. The proportion of females employed and retired is similar in each age-group in the NorStOP and the Staffordshire populations. However, the proportion of males employed in the NorStOP is lower than in Staffordshire, with a higher proportion retired in each age-group.

The lower level of employment in the NorStOP could be due to the decline in local industries in recent years, for example mining and pottery making, because the industry of North Staffordshire is not necessarily well represented by the industry of Staffordshire as a whole. Alternatively, it could be that those who do not work have more time to complete such surveys and were more likely to respond to the baseline NorStOP questionnaire. Although this appears to be the case only in males, it could be a form of response bias, which will be discussed further in Section 4.5.2. Employment could also be lower in the NorStOP than in England overall as a result of poorer health in the NorStOP population. This latter reason is a greater cause for concern as it is likely to influence the prevalence and potentially the severity of locomotor disability in the population. Again this will be discussed in Section 4.5.2.

4.4.2.7 Occupational class

The Staffordshire population has slightly more people in lower occupational classes and fewer people in higher classes than in England as a whole (Table 4.9). This difference is more marked still in NorStOP cohort.

The differences between the NorStOP and England are likely to be a result of employment opportunities in the North Staffordshire area, which have traditionally been in manual occupations. As lower occupational class has been shown to be related to poorer health, the larger numbers of people in lower occupational classes in the NorStOP may result in a higher level of locomotor disability than would be found in England as a whole. However, this should not detract from the value of analyses within the dataset.

Table 4.8 Distribution of the employment status in the English and Staffordshire populations and the NorStOP responding cohort, by gender and age-group

	England[a]		Staffordshire[a]		NorStOP responders[b]	
	Employed (%)	Retired (%)	Employed (%)	Retired (%)	Employed (%)	Retired (%)
Males						
50 to 54 years	81.2	2.6	81.7	2.5	75.8	3.3
55 to 59 years	71.5	8.2	71.4	7.9	65.5	9.7
60 to 64 years	49.1	24.3	47.3	24.2	42.6	26.6
65 to 69 years	14.6	78.0	11.8	80.7	6.6	88.8
70 to 74 years	7.2	86.0	5.9	86.6	1.8	97.7
Females						
50 to 54 years	69.6	2.5	68.6	2.7	70.3	2.3
55 to 59 years	55.8	10.0	52.9	10.6	53.1	6.3
60 to 64 years	25.5	61.6	22.7	64.8	20.6	64.2
65 to 69 years	8.4	82.0	6.6	84.0	4.6	86.4
70 to 74 years	3.7	86.0	3.1	86.3	2.1	89.4

[a]Calculated from 2001 Census: standard tables downloaded from Nomis https://www.nomisweb.co.uk/home/Census2001.asp, 7 March 2008; [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire

4.4.2.7 Adequacy of income

The measure of adequacy of income used in the NorStOP was suggested by Thomas (1999) and was tested in the piloting of the NorStOP study at baseline, where it showed good reproducibility over a four-week period (Ross Wilkie, personal communication, 20 September 2010). However, no other studies, either nationally or regionally, could be found that had used this assessment of perceived income adequacy. It is therefore not possible to compare the NorStOP sample to the English population or any other sample, based on this item.

In a study in Leicestershire in 1988, Matthews et al (2005) showed that 81% of those aged 75 years and over found their income to be adequate using the question "*Do you find this [your income] adequate or is it difficult to manage?*". Table 4.10 shows the distribution of response to the item in the NorStOP dichotomised as 'inadequate' (find it a strain to get by from week to week, have to be careful with money) and 'adequate' (able to manage without much difficulty, quite comfortably off). More than half of people perceive their income to be adequate and there is a general trend for this proportion be higher at older ages. However, levels of perceived adequacy are substantially lower in the NorStOP than in the study of Matthews et al (2005).

Table 4.9 Distribution of occupational class in the English and Staffordshire populations and the NorStOP responding cohort, by age-group

Age-group (years)	England			Staffordshire			NorStOP responders[b]		
Occupational class[a]	50 to 59 (%)	60 to 69 (%)	70 to 74 (%)	50 to 59 (%)	60 to 69 (%)	70 to 74 (%)	50 to 59 (%)	60 to 69 (%)	70 to 74 (%)
1	16.1	14.1	18.9	9.3	10.0	9.9	6.2	5.9	5.3
2	24.0	19.7	17.5	22.8	22.4	16.3	15.2	12.1	11.0
3	11.2	10.0	7.9	10.1	10.5	7.3	11.2	12.1	12.2
4	12.3	14.6	24.5	12.4	17.6	29.4	5.9	6.6	6.5
5	8.9	9.1	5.3	11.0	13.2	6.8	5.5	6.8	7.1
6	15.3	16.8	12.4	17.6	21.5	12.4	28.2	23.8	22.8
7	12.2	15.7	13.5	16.8	24.9	18.0	27.9	32.8	35.1

[a]Classifications derived from Office for National Statistics (2000 and 2002). 1 Higher managerial and professional, 2 Lower managerial/professional, 3 Intermediate occupations, 4 Self-employed, 5 Lower supervisory/technical, 6 Semi-routine, 7 Routine; [b]Age-group at baseline in those responding to the baseline Health Survey questionnaire

Table 4.10 Distribution of adequate income in the NorStOP responding cohort, by age-group

Age-group[a]	% with adequate income
50 to 54 years	57.2
55 to 59 years	55.5
60 to 64 years	53.4
65 to 69 years	57.0
70 to 74 years	54.6
75 to 79 years	53.4
80 to 84 years	55.8
85 to 89 years	58.2
90 years and over	64.8

[a]Age-group at baseline in those responding to the baseline Health Survey questionnaire

4.4.3 Attrition from the NorStOP1 over the six years of follow-up

At the six-year follow-up of the NorStOP1, 25% (n=2,831) of the original mailed sample remained in the study (Table 4.11). There were similar rates of response in males and females over the course of the follow-up. Over time, there was a higher rate of response in those aged under 70 years than in those age 70 years and over. This is potentially due to the ageing of the population and higher death rates in the older age-groups. Those with further education, non-manual occupational class and perceiving their income to be adequate were more likely to remain in the study than those with school-age education only, manual occupational class or perceiving their income to be inadequate. Those remaining in the study tended to have higher scores on both the physical and mental health components score of the SF-12 (Ware et al 1996) at baseline, indicating better functioning.

4.5 Suitability of the NorStOP for use in this thesis

This section addresses the suitability of the NorStOP datasets to meet the objectives of this thesis. Briefly, these objectives were:

- to assess current approaches to the measurement of locomotor disability in a general population and their suitability for this purpose;

- to consider the suitability of the PF-10 for this measurement purpose, in terms of the scale as a whole and individual items, and to develop and test an interval-level measure of locomotor disability using Rasch analysis;
- to demonstrate the use of the individual PF-10 items and an interval-level measure in investigating certain aspects of the general population epidemiology of locomotor disability.

In order to meet these objectives, data used in this thesis should be as representative as possible of a general population aged 50 years and over. This section sums up the evidence presented in this chapter around the NorStOP as an example of such a general population and discusses this datasets' usefulness for this PhD project.

Table 4.11 Attrition from the NorStOP1 over the six years of follow-up

	Mailed at baseline	Responded at baseline	Responded at three-year follow-up	Responded at six-year follow-up
Overall[a]	11,309 (100)	7,878 (70)	4,234 (54)	2,831 (67)
Gender				
Male	5,116 (45)	3,462 (44)	1,884 (45)	1,233 (44)
Female	6,193 (55)	4,416 (56)	2,350 (56)	1,598 (56)
Age-group				
50 to 59 years	3,871 (34)	2,521(32)	1,566 (37)	1,151 (41)
60 to 69 years	3,177 (28)	2,352 (30)	1,413 (33)	1,025 (36)
70 to 79 years	2,689 (24)	2,030 (26)	971 (23)	544 (19)
80 years and over	1,570 (14)	975 (12)	284 (7)	111 (4)
Educational attainment[b]				
Further education	-	823 (11)	525 (13)	404 (15)
School-age education only		6,848 (89)	3,626 (87)	2,377 (85)
Occupational class[b]				
Non-manual	-	2,119 (31)	1,397 (37)	1,029 (41)
Manual		4,619 (69)	2,342 (63)	1,506 (59)
Perceived adequacy of income[b]				
Adequate	-	4,177 (55)	2,389 (57)	1,682 (60)
Inadequate		3,481 (45)	1,778 (43)	1,111 (40)
SF-12 score at baseline[b,c]				
Physical component summary	-	40.7 (12.5)	42.0 (12.4)	43.2 (12.1)
Mental component summary	-	48.8 (11.2)	49.6 (11.1)	50.1 (10.9)

Numbers are n (%) unless otherwise stated; [a]Percentage is of responders from n in column immediately to the left; [b]Available only for those individuals responding at baseline; [c]Mean (SD)

4.5.1 Recruitment in to the study

The sampling frame from which the NorStOP cohort was recruited was the list of people registered with eight general practices in the local area. Approximately 98% of the British population are registered with a general practice (Bowling 2002), making such a list suitable for this type of study. It is known that practice lists may include out-of-date information regarding address or vital status, as much of the time the updating of records, particularly regarding address changes, relies on the practice being provided with this information by the patient. Despite this, a practice register is one of the most viable methods of recruiting participants for studies such as the NorStOP. Furthermore, the use of the NHSSTS reduced the problem of missing or inaccurate data at follow-ups.

The practices involved in the NorStOP were chosen for two main reasons i) the large size of their registered populations made the administration of the survey easier, and ii) they were not involved in other similar research at the time of the baseline survey. For the NorStOP2 and NorStOP3 subcohorts, proximity to the local hospital was an additional consideration as people were to be invited to clinical assessments for the CAS-K and the CAS-HA. Hence, the practices were not randomly chosen from the local area, but were selected based on practical criteria. Everyone aged 50 years and over was eligible to take part in the study, unless excluded by their GP, and there is no reason to suspect that the people registered with the chosen practices were systematically different from the local population. The mailed sample was however older than the population of Staffordshire as a whole; no data were available specifically on the population of North Staffordshire.

The sample was not therefore randomly chosen from the whole UK older adult population, as would be desirable to fully generalise the findings to the UK. However, the local sampling frame had similar demographic characteristics as the regional and national populations, but appears to be dissimilar on ethnic and employment factors. The findings from the NorStOP sample should be generalised only with great caution to the national population.

4.5.2 Response bias

Having established that the intended sample for the NorStOP presented a useful sample for this thesis, it was necessary to ensure that a good proportion of the sample took part in the study

and that those taking part were representative of those contacted, and hence of the underlying population.

In order to achieve a good response rate in the NorStOP, a repeat mailing strategy was used. This method of survey administration had previously been used in studies at the Research Centre and proved successful (Jinks et al 2001, Boardman et al 2003). Bowling (2002) suggested that response rates of 75% and above could be considered good. The response to the NorStOP baseline Health Survey questionnaire, although slightly lower than this at 71%, was still reasonable.

In the NorStOP, information on age and gender was available from the general practice lists for all those mailed. This is unusual and allowed some evaluation of potential biases in response to the postal survey. The analyses in this chapter have shown that non-response was higher at older ages in females whilst non-response was highest in the youngest age-group for males. However, there was no trend for those responding to the baseline NorStOP Health Survey questionnaire to be generally younger or older than those in the English population.

Overall, there is some evidence that males, those who live alone and those in lower socioeconomic groups might be underrepresented in the NorStOP at baseline compared to the English and Staffordshire populations. However, it is not clear if these differences in the makeup of the NorStOP and more general populations are a result of bias in the response to the NorStOP, or of a difference in the definitions used to derive population subgroups.

4.5.3 Follow-up

The NorStOP is a cohort study. That is, the same people are followed over a period of time (three time points over a total of six years in the case of the NorStOP1 data to be used in this PhD). This is useful in that the temporal patterns of locomotor disability can be investigated, as required by Objective 5 (Chapter 1).

As when recruiting participants into a study, it is necessary to maintain a high participation rate when following up a cohort. The NorStOP employed the same repeat mailing strategy at follow-up mailings as at baseline in order to achieve as high a response rate as possible. This strategy was successful to a large extent, but the cohort suffered much attrition when people refused permission to be recontacted after each follow-up.

4.5.4 Attrition bias

Similarly to recruitment into a study, it is important that those people who continue to be part of a cohort over time are representative of the original population from which they were recruited. Should those who continue to take part differ from those who do not, the data may suffer from response bias.

In this thesis, data from the follow-ups of the NorStOP2 and 3 subcohorts were not included because of the respective timings of these subcohorts and this thesis. However, the three NorStOP subcohorts were very similar in their age, gender and socioeconomic structure at baseline. Therefore, there is no reason to suspect that the use of only the NorStOP1 subcohort over the six years of follow-up should introduce bias into the longitudinal elements of the analyses in this thesis (assessment of responsiveness in Chapters 6 and 12, and the longitudinal course of locomotor disability in Chapter 14).

However, on considering loss to follow-up within the NorStOP1, there was considerably more loss to follow-up in those who were older, in lower socioeconomic groups and who had worse physical and mental health at baseline. Although this differential loss to follow-up is almost inevitable in a sample of older people such as the NorStOP1, this is likely to result in a biased sample over time.

In particular, the sample on average became younger, as older people were differentially lost to follow-up, and healthier (in terms of baseline reported health), as those in the worst health are lost from the sample. In the case of the NorStOP1, those from lower socioeconomic groups, who are generally known to report worse health (Department of Health 2003), were also lost to follow-up at a higher rate than would be expected, creating further potential biases. This is always an issue in longitudinal studies, and it is necessary to recognise the limitations this imposes on a study. In this thesis, these issues are considered as they affect the analysis presented in the following chapters. However, there is a wider issue about how attrition bias affects studies in general. For example, many of the longitudinal studies identified in the literature review in Chapter 3 are likely to have been affected by similar issues of attrition, which may have underestimated the onset and progression of locomotor disability and overestimated recovery rates.

4.5.5 Data quality

The NorStOP surveys were administered by post, which was the only practical option given the size of the study. This does mean though that there was no control over who completed the questionnaire. Basic information such as gender, date of birth, name and address information from the general practice list were checked against completed questionnaires and consent forms to ensure that the intended person responded. However, some people did not complete these items on the questionnaires; in this instance, no checks could be made. Even where gender and date of birth corresponded to the database, it is possible that another person significantly influenced the responses given. This may especially have been the case where the intended respondent could not read the questionnaire themselves, either because of a language barrier, a low level of literacy or an eyesight problem. Given the high proportion of white British people in the local population, it is unlikely that many people could not complete the survey because of a language barrier and school-attendance was compulsory in this age-group, suggesting that illiteracy levels should be minimal. However, it is not possible to gauge the extent to which this older adult population might have been affected by sight problems.

In order to efficiently convert the data on the paper questionnaires into a computerised format for analysis, the data were scanned and processed using the Tele*form*™ software package (Cardiff Software Inc 1998) (with the exception of the pilot study which was manually entered into a Microsoft Access database). The pain manikins were entered manually, using a transparent overlay to define pain areas (Figure 4.1), into a Microsoft Access database. One in ten of the questionnaires were manually checked for data entry errors. Further to these checks, an independent member of the research team checked that all data had been entered for all respondents and double checked that dates of birth matched across databases.

It is reassuring that every effort was made within the NorStOP to ensure that data was of as a high a quality as possible. There were however some instances of high levels of missing data within some constructs used in the analyses in this thesis (Table 4.1). This was particularly clear where multiple items were required in order to calculate a score for a construct, for example the SF-12 physical and mental health component scores. These higher levels of missing data on some items will reduce the sample size available for analysis below that suggested by the overall response rate to the survey. Furthermore, if subject to bias in terms of which participants failed to provide data for a particular construct, this could introduce further bias in terms of response and

attrition, as discussed above. It is not possible to clearly ascertain whether those people missing items are in worse health than those who complete them, although this would seem plausible.

4.6 Other potential sources of data

This PhD project concentrates on data from the NorStOP cohort, with some use of other, external cohorts (Keele Knee Pain Cohort Study (Jinks et al 2004), Welsh Health Survey (National Assembly for Wales 2000) and a Dutch cohort (van der Windt et al 2008)) in Chapter 11. This begs the question of whether another dataset would have been a more suitable setting in which to conduct this PhD project.

There are three key areas to consider in the choice of a suitable dataset for this thesis: the generalisability of the sample, the presence of suitable items with which to measure locomotor disability and test the appropriateness of this measure, and the presence of longitudinal data in order to fully demonstrate the potential of any new measure of locomotor disability.

In terms of the representativeness of the sample, and the potential to generalise the findings presented in this thesis, some of the publically available data sets such as the Health Survey for England (for example, Department of Health 2005), or even the national Census may appear more appropriate. These studies rarely collect detailed data on disability. For example, in 1996, the Health Survey for England included the PF-10, but only for those people under the age of 20 years. Furthermore, although these nationally representative surveys are carried out on a regular basis, they do not generally provide the opportunity to follow individuals over time. In the case of the Health Survey for England, a representative sample of the population is taken at each wave of the survey.

A publically available and nationally representative dataset that collected more detailed information on health and disability, and which is following participants longitudinally is the English Longitudinal Study of Ageing (ELSA). The ELSA is based on a sample of people aged 50 years and over, who had previously taken part in the Health Survey for England. To date, these people have provided information on four separate occasions. Whilst the ELSA is a very useful resource and data from the first two waves are publically available, it does not contain the same level of detailed information on disability and health as the NorStOP. This detail was needed in order to fully develop and test a measure of locomotor disability.

The SF-36 is the world's most widely used health assessment instrument (Garratt et al 2002), and as reported in Chapter 3, its physical functioning subscale, the PF-10 has been suggested by some authors as a potential measure of locomotor disability (for example Bohannon et al (2004), Peat et al (2006a)). The PF-10 appears to be the strongest candidate measure for locomotor disability among the plethora of health assessment instruments available (Feinstein et al 1986, de Vet et al 2003). It is essential then that any study used in this PhD project contains this instrument. Due to its length, the SF-36 is rarely included in general surveys, such as the Health Survey for England or ELSA; often being confined to more specific studies, such as the NorStOP.

Given the availability of the NorStOP and the detailed level of information collected on the same participants over a long period of time, this was considered to be the most appropriate source of data for use in this PhD project.

4.7 Conclusions

This chapter has considered the North Staffordshire Osteoarthritis Project as a potential source of data for use in addressing the objectives of this thesis. There is evidence that although broadly representative of the local population, the NorStOP is not fully representative of the national population in terms of socioeconomic status and ethnicity. This seems to be partly due to the local population structure in North Staffordshire being different to that of the England as a whole, and to some degree due to response bias. In terms of the follow-up of the NorStOP over time, there is obvious potential for bias in longitudinal analyses due to selective loss to follow-up. The causes of this loss to follow-up and of levels of missing data, both cross-sectionally and longitudinally are not clear, and their effects will differ in different analyses. These effects and their potential to significantly bias the findings of the various analyses presented in this thesis will be discussed as they arise in the following chapters.

Despite the obvious limitations of the NorStOP, it still presents a useful resource for use in this PhD project. In particular the NorStOP1 subcohort has longitudinal data covering a six-year period for use in longitudinal analyses. Crucially, participants in the NorStOP were asked to complete the PF-10, which has been suggested as a potential measure of locomotor disability (for example Bohannon et al (2004), Peat et al (2006a)), at every time point. Chapter 6 considers the evidence around this suggestion, both in the current literature and in the NorStOP datasets, where

additional analyses are conducted. First, Chapter 5 describes the psychometric properties required of a measurement instrument.

5 The required psychometric properties of measurement instruments

5.1 Introduction

Chapter 2 defined measurement, for the purpose of this thesis, to be "*the use of an interval- or ratio-level scale to assess an underlying construct*". Many instruments have been developed in order to measure a variety of constructs in health research and practice. Some of these instruments have been more successful in meeting their intended measurement aims than others. In order to differentiate between these instruments, in terms of their quality, several sets of criteria (Bombardier and Tugwell 1987, Lohr et al 1996, Bot et al 2004, Terwee et al 2007) have been developed that assess key psychometric properties of the instruments. Many aspects of these quality criteria are very similar.

This chapter describes the psychometric criteria that will be applied in this thesis in order to fully evaluate the measurement of locomotor disability, both in terms of the current state of measurement, and the assessment of newly proposed ideas. The criteria used in this thesis are based primarily on those proposed by Terwee et al (2007), as this is the most recent and comprehensive set of standards. However, it will also draws on some of the ideas from Lohr et al (1996) and Bot et al (2004) that are not covered by Terwee et al (2007).

5.2 Key psychometric properties described in the literature

The majority of the published criteria for assessing the psychometric properties of measurement instruments are designed for use in comparing instruments (Lohr et al 1996, Bot et al 2004, Terwee et al 2007). This comparison would usually be necessary in order to choose the most suitable outcome to include in a study. Hence these criteria often aim to assign a positive or negative rating to each criterion. Although the next chapter assesses the suitability of one potential measure of locomotor disability – the PF-10 – on the whole, psychometric criteria are used in this thesis to guide the development of a new measure of locomotor disability. The assignment of positive and negative ratings is not then wholly appropriate.

This section describes each of the measurement properties suggested in previous sets of criteria, before they are summarised as they will be used in this thesis.

5.2.1 Conceptual and measurement model

Lohr et al (1996) discussed, what they referred to as, the "*conceptual and measurement model*" of a scale. This refers to the background to the scale and the concepts that it covers. Within this criterion, one should consider the scale of measurement and why it was chosen, as well as how items are related and combined. This criterion also incorporates procedures to handle missing data.

5.2.2 Repeatability

Repeatability refers to the ability of an assessment instrument to produce results that can be replicated. This is a particularly important property of these instruments because it is necessary (although not sufficient) in order for the scale to be valid: an instrument cannot be assessing what it is supposed to assess if it is not producing a consistent score. As a result, repeatability is included as a concept in most checklists (Bot et al 2004, Lohr et al 1996, Terwee et al 2007), although a range of synonyms are used to describe it: reproducibility, reliability, agreement.

According to the framework of Terwee et al (2007), reproducibility can be split into two distinct areas: agreement, i.e. the similarity of repeated measurements or the size of measurement error, and reliability, i.e. ability to distinguish individuals from one another (de Vet et al 2006).

5.2.2.1 Agreement

Agreement is the closeness of scores on repeated occasions (for either different raters or different time points); in other words a lack of error, either systematic or random (Terwee et al 2007). In this thesis, interest is in self-reported measures and so agreement between raters is not relevant. Agreement over time is important in scores that will be used for evaluative purposes (de Vet et al 2006), i.e. to follow individuals over time.

Agreement is often described using the standard error of measurement[2] (SEM), or by comparing the smallest detectable difference[3] (SDD) to the minimal important change (MIC) (see below) (Terwee et al 2007). Due to the mathematical operations required to produce these

---

[2] $SEM_{agreement} = s\sqrt{1 - ICC_{agreement}}$ where *s* pooled standard deviation of scale scores at the test and retest and $ICC_{agreement}$ is the intraclass correlation coefficient for agreement (see Section 5.2.2.2)

statistics, they are only appropriate when the scale score is on an interval-level. In the case of ordinal scales, agreement can be assessed using Kappa statistics (Bot et al 2004), which quantify the agreement between two ordinal-level scales beyond that of chance. Kappa statistics can be weighted in order to give more influence to values at the second time point that are closest to the values at the first time point, and confidence intervals can be calculated using bootstrapping techniques (Reichenheim 2004). The number of bootstrap samples required is usually dependent on the desired precision of the estimate. Altman (1991: pg 404) provided criteria against which to judge the strength of agreement using Kappa statistics.

5.2.2.2 Reliability

Reliability is the extent to which individuals can be distinguished from each other despite measurement error (de Vet et al 2006). That is, is the variation between individuals sufficiently larger than the 'noise' in the scale score within individuals? Hence, reliability is dependent on the heterogeneity of the population sample and as such is reduced in more homogeneous samples. Reliability is important when considering the instrument for use in distinguishing between individuals (de Vet et al 2006): the actual score assigned to the individual on repeated measurements is unimportant, provided that the rank ordering is the same and the measurement error is sufficiently small in relation to the heterogeneity in the population.

Intraclass correlation coefficients are recommended for the quantification of reliability (Terwee et al 2007). Several forms of this coefficient are available (Streiner & Norman 2003: pg 133-137). Pearson's correlation coefficient should be avoided in the assessment of reliability. Due to the calculations required to calculate an intraclass correlation coefficient, this method of assessing reliability is only suitable for use with interval-level scale scores.

5.2.3 Unidimensionality

If the items within an instrument are to be combined into a scale, it is important that they are all related to the same underlying concept, so that any change in the scale score can be attributed to that concept. If a scale consists of items from dimensions other than the one of specific interest, then a change in scale scored cannot be assumed to reflect a change in the concept of interest. No

---

[3] $SDD = 1.96SD_{change}/\sqrt{n}$

published checklist uses the term unidimensionality, often referring instead in 'internal consistency' (Bot et al 2004, Terwee et al 2007). For the purpose of this thesis, it was considered that 'unidimensionality' clearly and concisely described the concept of interest, and so this term will be adopted throughout.

The most common method for the assessment of unidimensionality is factor analysis. Factor analysis techniques can be categorised into two major forms: exploratory and confirmatory. Exploratory factor analysis attempts to find the factor structure that provides the best fit for the data, whilst confirmatory factor analysis assesses whether the fit of the data to a hypothesised factor structure is acceptable. It is commonly acknowledged that exploratory factor analysis should be used when one has a set of items and does not know the appropriate factor structure that is appropriate; whilst confirmatory factor analysis should be used when a factor structure has been hypothesised.

Another common method of assessing unidimensionality is to assess a concept known as internal consistency. This is the extent to which all the items in a scale are assessing the same concept (Terwee et al 2007). To ensure this is the case, all items should be moderately correlated with each other and each item should correlate with the overall scale score. However, if the correlation between items is too high, some of the items are likely to be redundant, as they are assessing very similar parts of the construct.

One way of testing the internal consistency of items is to consider the correlation between each item and the scale score calculated without that item, i.e. item-total correlation. Streiner and Norman (2003) suggest that the correlation between each item and the scale score without that item should be at least 0.2. Another method of assessing internal consistency is to use 'split-half' reliability. That is, the scale is randomly split into two and the new 'subscales' correlated with each other. This method generally leads to an underestimate of the internal consistency because the two 'subscales' are shorter than the original scale. The Spearman-Brown prophesy formula can allow for this by taking account of the factor by which the scale has been shortened, k, and the correlation, r, between the original subscales: $\frac{kr}{1+(k-1)r}$ (Streiner & Norman 2003). Also, there are many ways that the scale can be split into two. To overcome this problem, the most common way of assessing internal consistency is to use Cronbach's alpha statistic, which essentially averages over all the possible split-half reliability statistics, to give an overall picture (Streiner & Norman

2003). There are however problems with using Cronbach's alpha (Cronbach 1951). First, the value of alpha will increase as the number of items in the scale increases, regardless of the correlation between items (Streiner & Norman 2003). Second, if items from two scales that are correlated are mixed together in one scale, alpha will be high, despite the fact that the items are not homogeneous (Streiner & Norman 2003). It is therefore necessary to use principle components or factor analysis to ensure that the items form a single scale, before calculating alpha. A value of alpha greater than 0.7 can be considered to indicate good internal consistency, whilst a very high value of alpha (>0.9) could suggest redundant items. That is, two or more items are tapping into the same part of the construct and one or more of them could be removed to simplify the tool (Streiner & Norman 2003).

5.2.4 Validity

Testing the validity of a scale is concerned with ensuring that the scale is assessing the concept that it was intended to assess. As such, it is included in all checklists and textbooks on the subject (Bombardier & Tugwell 1987, Bot et al 2004, Lohr et al 1996, Streiner & Norman 2003, Terwee et al 2007), although it is often referred to in different ways. For the purpose of this thesis, there are considered to be four types of validity: face, content, construct and criterion. These four types are often described as representing levels of evidence, and relate to the rigour with which the validity is tested; face validity providing the lowest level of evidence and criterion validity the highest.

Also to be considered in terms of validity, are the interpretation of the scale, i.e. assigning qualitative meaning to scores, the presence of floor and ceiling effects, and levels of responsiveness or sensitivity to change. The following sections describe each of these aspects of validity in turn.

5.2.4.1 Face validity

Face validity is simply an assessment of whether an instrument appears 'on the face of it' to be assessing what it is supposed to assess (Streiner & Norman 2003). Usually it is expert opinion that decides whether there is face validity, although it is often argued that face validity is important in gaining the cooperation of participants (Streiner & Norman 2003) and so their views should also be taken into account.

In some situations, face validity is thought to be a negative characteristic (Streiner & Norman 2003), because social desirability might discourage people from answering truthfully, for example in a questionnaire about excessive drinking. It can also be undesirable if the researcher suspects that the respondent may wish to 'please' the questioner.

5.2.4.2 Content validity

Content validity is closely related to face validity in that it is subjective. However, content validity goes further, looking at whether the instrument is covering everything within the desired concept and nothing outside it. Streiner and Norman (2003) describe these two strands of content validity as content coverage (ensuring people do not differ according to some aspect of the construct that is not assessed) and content relevance (ensuring that the differences between people are all relevant to the concept being measured). These authors suggest that the relative importance of different parts of the construct should be displayed in the instrument: more items should represent those concepts that are thought to be more important. Content validity, although sometimes ignored (Pollard et al 2007), is important as a scale with high content validity allows broader inferences to be made from calculated scores.

In addition to these traditional criteria for content validity, Terwee et al (2007) demand that the aims of the assessment instrument, i.e. what it is attempting to assess and why, the target population and concepts to be covered are clearly stated. They also require that patients or participants, as well as experts, were involved in item selection before content validity can be accepted.

5.2.4.3 Construct validity

Terwee et al (2007) describe construct validity as, "*the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured*". As such, it could be considered a stronger form of evidence regarding validity than face or content validity.

To assess construct validity, specific hypotheses are generated about how the scores on the instrument being tested will differ in different population groups, for example males versus females, or about how the scores will correlate with scores on other instruments (Streiner & Norman 2003). This may be a fairly high correlation (the instruments are assessing similar constructs), or a low to

moderate correlation (the instruments are assessing related but not particularly similar constructs). Usually a very high correlation with another tool would not be desirable as this would mean the new instrument is extremely similar to the old one, and possibly redundant. Ideally, the suggested hypotheses should be plausible and very specific, including the expected difference between groups or the level of correlation expected, not just p-values (Terwee et al 2007).

Although construct validity is intuitively simple, it does have some problems: if a hypothesis is upheld, then there is evidence towards construct validity. However, if the evidence is insufficient to reject the null hypothesis, for example, two groups are found to have similar scores on the instrument, then there are a number of possible explanations for this: the instrument being tested may not be valid, the suggested hypothesis may have been incorrect, the classification of the groups could have been wrong, or the analysis could be underpowered, i.e. a type II error. Establishing construct validity is then, an on-going process in which many hypotheses are formed and tested, eventually culminating in the acceptance that validity has or has not been shown.

Terwee et al (2007) suggested that to accept the construct validity of an instrument, hypotheses should be rejected in no more than 25% of cases and that each subgroup used in analyses should contain at least 50 individuals. They did not state how many hypotheses should be tested, nor how one is to judge that the hypotheses were correctly formulated, classification of groups accurate and the calculation suitably powered.

As has been mentioned for previous measurement properties, the level of measurement of the scales involved in testing construct validity should be taken into account. This applies equally to the scale score under investigation, and any scores against which it is being assessed.

5.2.4.4 Criterion validity

Criterion validity, in which an instrument is compared to a widely accepted ‘gold standard’ is often considered to be the highest level of evidence in assessing validity. Terwee et al (2007) suggest a correlation between the measure being tested and the gold standard of at least 0.7 is required. However, it is only possible to compare a tool with a gold standard when that standard exists. Where a standard is thought to exist, it is necessary to ensure that it is truly a gold standard and not just something that is widely used.

5.2.4.5 Floor and ceiling effects

Floor effects occur when a large proportion of respondents, Terwee et al (2007) suggest 15%, are assigned the score that reflects worst health on a scale. Similarly, ceiling effects occur when a large proportion are assigned the score that reflects the best possible health state (Vogt 1993). Floor and ceiling effects are caused when the distribution of the scale scores is not Normal. Often this is the result of a very high or low functioning sample or because the tool does not adequately sample the domain. It is necessary to detect floor and ceiling effects because they can indicate limited content validity and also reduce responsiveness and reliability. This is because it is not possible to detect a change in status within the floor or ceiling and those at the ends of the sale cannot be distinguished from each other, even though they may differ with respect to the construct of interest.

5.2.4.6 Interpretation

Health assessment instruments often yield a single figure or a profile of figures for each individual, but these values do not mean anything by themselves. In order to fully utilise such a scale, it must be possible to attach some sort of qualitative meaning to the scores (Bot et al 2004, Lohr et al 1996, Terwee et al 2007). Several ways of attaching this meaning have been suggested. The simplest method is to consider population normative data (norms): what does the distribution of scores look like in an average population?

To be a useful comparator, a sample for normative data should be randomly drawn from a representative sampling frame and biases minimised. The mode of survey administration (postal or interview) has been shown to affect responses, with interview-administered surveys being more prone to interviewer- and social-acceptability bias for example, resulting in more favourable average scores (Bowling et al 1999). It is therefore important to use norms for a similarly administered survey. Another issue surrounding norms is what the ‘normal’ population should be. Many studies are conducted regionally, but regional scores often do not reflect a national population, as local differences can occur (Bowling et al 1999).

In addition to presenting normative data, many authors suggest the calculation of a minimal important change (MIC) or minimal clinically important difference (MCID). These values represent the change over time or difference between groups, respectively, that can be considered clinically meaningful. There are various distribution- and anchor-based methods available for the calculation

of an MIC. A distribution-based approach relies on the statistical properties of the sample, and for this reason is often criticised (for example de Vet et al (2007) Terwee et al (2007)). Anchor-based approaches use an external criterion to define a change which can be 'mapped' to a change in the score. Terwee et al (2007) recommend an anchor-based approach to developing an MIC, whilst de Vet et al (2007) have developed the anchor-based distribution method which combines the two approaches. Whilst the use of an anchor may be methodologically more rigorous, and the preferred approach, such an anchor may not be available, making calculation of a MIC difficult.

Whatever method is chosen to provide information on the interpretation of the score, be it normative data or a MIC it should be suitable for the score's level of measurement. For example, it does not make sense to present means and standard deviations for ordinal-level scales or to calculate an MIC in the context of the response to an ordinal-level item response.

5.2.4.7 Responsiveness

Responsiveness, or sensitivity to change, is another key property that is included in the majority of checklists (Bombardier & Tugwell 1987, Bot et al 2004, Lohr et al 1996, Terwee et al 2007). Responsiveness can be considered to be longitudinal validity, as it examines whether real changes in the construct being assessed are reflected in changes in the scale score. Responsiveness is particularly important when it is intended to use the instrument to consider changes over time, as it is necessary to know that the instrument can detect these changes. Similarly to validity, responsiveness can be tested using hypotheses, either comparing the measure to a series of associated constructs, or to a gold standard if one exists. Analogous issues arise in the evaluation of responsiveness as in the evaluation of validity: hypotheses must be specific and specified a priori, whilst gold standards can be difficult to find (Terwee et al 2007), and statistical testing must be appropriate to the level of measurement of any scores used in the testing process.

5.2.5 Feasibility

In order for a health assessment instrument to be useful in practice, it must be feasible. That is, it must be acceptable to respondents, easy to use for the administrator, attract high levels of complete data and be relatively good value in terms of time and money. Many checklists include various aspects of feasibility (Bombardier & Tugwell 1987, Bot et al 2004, Lohr et al 1996). The

following sections describe the different aspects of instrument feasibility and how they can be tested.

5.2.5.1 Respondent burden

It is important that health assessment instruments do not place undue strain on the respondent (Lohr et al 1996), whether this is physical or emotional stress, or simply the inconvenience of having to find medical or financial records in order to answer the questions. Another important factor considered by many researchers is the time to complete the questionnaire. This is important in terms of the feasibility of the questionnaire, and in particular its suitability for inclusion in a booklet with other instruments. However, the time taken to complete the instrument will obviously fluctuate with the method of administration and the ability of the respondent. When instruments are to be used in a self-completion setting, for example in a postal survey, the reading age of the items must also be considered, as respondents must comprehend the questions.

5.2.5.2 Administrator burden

A health assessment instrument must be straight forward to use for the researcher or clinician, as well as for the respondent. Once the respondent has completed a questionnaire, it is necessary to collate the data and score the instrument. Complex formulae for scoring can make this process very difficult and require specialist knowledge. The cost of obtaining, administering and scoring the tool is also important, as resources are usually limited. A further consideration in some studies might be the availability of adaptations or translation of the instrument for use in different cultures or languages.

5.2.5.3 Missing data

Respondents often do not complete all of the items in an instrument, particularly in postal surveys. This can lead to reduced sample sizes and more importantly can be a form of bias, with respondents who miss items being different to those who complete them. Missing items can also be an indicator of the acceptability and comprehensibility: items that are repeatedly missed may be difficult to understand or may seem irrelevant to respondents. Missing data are usually reported in

the form of the percentage of respondents failing to provide usable data on each item. It is also usual to present the proportion of respondents for whom a scale score is available. This may be with or without the use of any mechanism for the imputation of missing item scores.

It is also prudent to consider the pattern of missing data. If there is a pattern to the missingness across items, this could indicate problems with certain items. A pattern of missing items responses in certain groups of respondents could indicate response bias.

5.3 Summary and discussion

When assessing the psychometric properties of a measurement instrument, there are many aspects to consider. Table 5.1 summarises these different aspects as they will be assessed in this thesis.

Several checklists have been proposed in recent years for the assessment of the psychometric properties of measurement instruments (Bombardier & Tugwell 1987, Lohr et al 1996, Bot et al 2004, Terwee et al 2007). Despite the range of checklists available, no single list presented a comprehensive set of psychometric criteria. This lack of a fully comprehensive list may be as a result of the purpose of these checklists, which was generally to compare measurement instruments, rather than to evaluate a single instrument, as is the case in this thesis (see Chapters 6 and 11).

A more comprehensive set of assessment criteria have recently been published (Mokkink et al 2010). These criteria are the results of a Delphi study that attempted to reach consensus on what a checklist of psychometric properties should include. Unfortunately, the publication of this paper was too late for it to be used as the basis for the assessment of psychometric properties in this thesis.

A caveat of the use of all the checklists published previously, and to an extent that of Mokkink et al (2010) too, is that the underlying concept of measurement on which they are based is not necessarily the same as has been defined in this thesis. These checklists do however provide a guide to the properties necessary for a good measurement instrument.

In the next chapter, the criteria discussed in this chapter, and summarised in Table 5.1, are used to assess the Short Form-36 Physical Functioning subscale (PF-10) as a potential measure of locomotor disability for use in epidemiological studies.

Table 5.1 Psychometric properties checklist used in this thesis

Property	Definition	Quality criteria
Conceptual and measurement model (Lohr et al 1996)	Background to instrument development and procedures used to create scale scores	Description and rationale for concepts Why particular scale of measurement , for example binary/ordinal Description of procedures used to create scales Scoring procedures including the handling of missing data
Unidimensionality (Bot et al 2004,Streiner & Norman 2003, Terwee et al 2007)	Extent to which the items in a scale form a single dimension	Suitable use of factor analysis to determine scales Item-total correlation >0.2 Cronbach's alpha ≥0.7
Repeatability		
Agreement (Bot et al 2004, Terwee et al 2007)	Absolute repeatability of a score over time	Use of limits of agreement, Kappa statistic, standard error of measurement
Reliability (Bot et al 2004, Lohr et al 1996, Terwee et al 2007)	Extent to which patients can be distinguished from each other despite measurement error	Intra-class correlation coefficient > 0.7 Correlation coefficients are inappropriate
Validity (Terwee et al 2007)		
Face (Streiner & Norman 2003)	On the face of it, the scale is measuring what it is intended to measure?	Comparison of contents to external markers of the construct
Content (Bot et al 2004, Lohr et al 1996, Streiner & Norman 2003, Terwee et al 2007)	Coverage of everything within the desired concept and nothing outside it	Content coverage and relevance Clear description of measurement aim, target population, concepts being measured and item selection Use of patient and expert panels in item selection
Construct (Bot et al 2004, Lohr et al 1996, Terwee et al 2007)	Ability of score to relate to similar scores and to differ between specified groups of people	Specific, sensible hypotheses about differences between groups or correlations with other scales are tested in a well-designed study At least 75% of hypotheses confirmed
Criterion (Lohr et al 1996, Terwee et al 2007)	Relation of the score to a gold standard	Convincing argument for a gold standard Correlation with gold standard >0.7
Floor and ceiling effects(Bot et al 2004, Terwee et al 2007)	>15% of scores indicating highest or lowest level of construct	Description of the proportion of responses in extreme categories

Interpretatability (Bot et al 2004, Lohr et al 1996, Terwee et al 2007)	Degree to which qualitative meaning can be assigned to scores	Normative data using appropriate summary statistics Relationship of scores to functional status or a clinical measure
Responsiveness (Bombardier & Tugwell 1987, Bot et al 2004, Lohr et al 1996, Terwee et al 2007)	Ability of score to change when the underlying construct changes	Specific, sensible hypotheses about differences between groups and at least 75% of hypotheses confirmed Comparison to a gold standard for which there is a convincing argument
Feasibility		
Respondent burden (Bot et al 2004, Lohr et al 1996)	Acceptable from respondents' point of view	Estimate of time taken to complete items No undue strain on respondent, for example emotional upset, the need for additional information on medicines Readability and comprehension levels tested and level found to be suitable for the population in which it to be used
Administrator burden (Bot et al 2004, Lohr et al 1996)	Acceptable from researchers' point of view Translations to other languages Cultural adaptations within a language	Estimate of cost to administer Description of method of scoring Description of facilities needed to administer/ calculate scores Translations and adaptations carried out appropriately
Missing data (Lohr et al 1996)	Levels of complete data should be high Missing data should be missing at random	Percentage of respondents not completing individual items Data are missing at random Percentage of completed scales and scales for which scores could be computed

Note that not all published checklists use the same terminology: concepts have been grouped under the terms used in this thesis

6 The suitability of the SF-36 Physical Functioning subscale (PF-10) as a measure of locomotor disability

6.1 Introduction

The Medical Outcomes Study Short Form-36 (Ware & Sherbourne 1992) is a widely used questionnaire designed to assess quality of life. It consists of 36 items, each mapping to one of eight subscales that assess various aspects of quality of life (Appendix C). One of these aspects is physical functioning and is assessed using the 10-item Physical Functioning subscale (PF-10). The PF-10 instrument is shown in full in Appendix B (Baseline Health Survey questionnaire, page 30, items a to j). As highlighted in Chapter 3, several studies (for example Bohannon et al (2004) and Peat et al (2006a)), have used the PF-10, or items from it, to assess locomotor disability, and Syddall et al (2009) asserted that it represents, "*a valid measure of mobility disability in epidemiological studies*". Should this be the case, it would prove very useful, as the SF-36 has been widely used in a variety of studies (Garratt et al 2002). Indeed, it has been translated and adapted for use in over 40 countries (Ware 2000), and these adaptations allow international comparisons.

Clear evidence of the suitability of the PF-10 score as a measure of locomotor disability would enable much more work in this area of disability research without the need to collect new data. During the course of this PhD project, Sydall et al (2009) produced an assessment of the internal consistency of the PF-10 and used dichotomised item and scale responses to predict performance in observed functional performance tasks. However, they did not conduct a detailed assessment of the PF-10 as a measure of locomotor disability. This chapter will carry out that assessment through analyses in the NorStOP datasets and consideration of the published literature.

6.2 The PF-10 as a measure of locomotor disability

As discussed in Chapter 1, there is no agreed upon definition of locomotor disability; in this thesis, it is defined as, "*the difficulty a person experiences in moving from place to place using the*

lower limbs" (Chapter 1). This section then, considers the conceptual and measurement model of the PF-10 as a potential measure of locomotor disability under this definition.

The PF-10 was designed to assess physical functioning, which for the purpose of the subscale's development was defined as, "*performance or capacity to perform a variety of physical activities normal for people in good health*" (Stewart & Kamberg 1992, cited in Mallinson 2002). Clearly this definition has a wider scope than the definition of locomotor disability used in this thesis. This leads into the measurement property usually referred to as content validity, i.e. whether a scale is covering everything within the desired concept and nothing outside it.

Cieza et al (2002) mapped the PF-10 items onto domains of the ICF (WHO 2002). They found that the Items a (vigorous activity) and b (moderate activity) did not map to any domain of the classification. The remaining items mapped to the "*activities and participation*" domain. Within this domain, Item j (bath/dress) was found to map to Chapter 5 (self-care), and so this item should not be considered an indicator of locomotor disability. The remaining seven items were mapped to Chapter 4 (mobility). However, the ICF definition of mobility, "*moving by changing body position or location or by transferring from one place to another, by carrying, moving or manipulating objects, by walking, running or climbing, and by using various forms of transportation*" (WHO 2002) is wider than the definition of locomotor disability employed in this thesis. Item c (lift/carry groceries) was mapped to the subchapter relating to "*carrying in the hands*". This is outside of the scope of locomotor activities as defined in this thesis. Item f (bend/kneel/stoop) was mapped to two subchapters of the ICF: "*getting into and out of a position*", which does not relate to moving from place to place, and to "*tilting the back downwards or to the side"*, which does not relate to the lower limbs. Of the remaining five items, two (Items d (climb several flights of stairs) and e (climb one flight of stairs)) were mapped to the subchapter relating to, "*moving the whole body from one place to another by means other than walking*". In particular, they mapped to the section of this subchapter relating to climbing, which is defined as "*moving the whole body upwards or downwards, over surfaces or objects, such as climbing steps, rocks, ladders or stairs, curbs or other objects*". Although this section of the ICF potentially covers a large range of tasks, it does relate to moving the whole body up or down, which was considered to include moving the body from one place to another. For this reason, Items d (climb several flights of stairs) and e (climb one flight of stairs) were considered to assess to locomotor disability as it is defined in this thesis. The three remaining items of the PF-10 (Items g - walk more than a mile, h - walk half a mile and i -

walk 100 yards) related to the subchapter of the ICF on walking, i.e. "*Moving along a surface on foot, step by step, so that one foot is always on the ground*" and in particular to the sections on walking long (more than a kilometre) and short (less than a kilometre) distances. Therefore these items were also considered to assess locomotor disability as defined in this thesis. Box 6.1 summarises which PF-10 items were considered to assess locomotor disability for the purpose of this thesis.

Box 6.1 Assessment of locomotor disability by the individual items of the PF-10

Item	Assessment of locomotor disability under definition in this thesis
a (vigorous activity)	No
b (moderate activity)	No
c (lift/carry groceries)	No
d (climb several flights of stairs)	Yes
e (climb one flight of stairs)	Yes
f (bend/kneel/stoop)	No
g (walk more than a mile)	Yes
h (walk half a mile)	Yes
i (walk 100 yards)	Yes
j (bath/dress)	No

Whether these five locomotor disability-specific items cover the full concept of locomotor disability is questionable. As shown in Table 3.1, some studies have included a greater range of items in their definition of locomotor disability, for example balance (Ebrahim et al 2000) and bending and straightening (Martin et al 1988). However, all studies included walking tasks, which are well represented in these five items from the PF-10, and the majority also included stair-climbing. This suggests that these five items from the PF-10 represent the core activities of the ill-defined concept of locomotor disability.

It seems logical that in order to use the PF-10 as a measure of locomotor disability, the five items not specific to locomotor disability should be removed, resulting in a five-item scale that is conceptually unidimensional, i.e. the items all map to the same concept. The next sections investigate these five locomotor disability-specific items as individual indicators of locomotor disability and then their possible combination into a measure using the same methodology employed to create the full PF-10 scale score.

6.3 Psychometric properties of the individual locomotor disability-specific PF-10 items

Having identified individual items in the PF-10 whose content is consistent with the concept of locomotor disability in this thesis, an evaluation of their psychometric properties is necessary. As well as using the PF-10 scale score created from a simple summation of the item responses (for example Peat et al (2006a)), previous studies have used individual items as indicators of locomotor disability (for example Bohannon et al (2004)). This section therefore considers the psychometric properties of the individual locomotor disability-specific items of the PF-10. It should be noted, however, that while these approaches are commonly encountered in the medical literature, neither the original summed rating score of the PF-10, or the individual item responses are consistent with the principles of measurement as outlined in Chapter 2.

6.3.1 Methods

Due to the nature of the individual items and the choice of these items for their conceptual relation to locomotor disability, some of the psychometric properties described in Table 5.1 were not considered relevant to these analyses. Therefore, this section will not consider the conceptual and measurement model, unidimensionality, reliability, face or content validity, floor or ceiling effects, interpretability or responsiveness. The evaluation of the remainder of the psychometric properties described in Table 5.1 is carried out via new analyses in the NorStOP datasets. Some sections also draw on data published by others.

6.3.1.1 Agreement

As the responses to the PF-10 item are on three-point ordinal-level scales, the most appropriate method with which to assess the agreement of the item responses over time is a weighted Kappa statistic. Quadratic weights (for example, Sim & Wright 2005) were used to assess agreement within the individual items between the test and retest administrations of the NorStOP pilot study questionnaire (Chapter 4). Analyses were carried out only in those respondents who reported the same level of general health on the SF-12 general health item on both questionnaires to ensure a real change in the level of functioning had not occurred. Confidence intervals were

calculated using 1,000 bootstrap samples, to give an estimate stable to three decimal places. Kappa statistics were interpreted using the criteria of Altman (1991: pg 404).

6.3.1.2 Construct validity

The construct validity each of the five items was assessed by forming hypotheses (Box 6.2), based on a search of the literature (Chapter 3).

BMI was categorised as described in Box 6.2 in line with guidelines from the Department for Health and Human Services (2007). In Hypothesis 2 (Box 6.2) the KAP items were considered dichotomously according to whether participantion restriction was reported in no domains or one or more domains. In Hypothesis 4, it was considered in terms of the number of restricted domains. The majority of people reported no or few domains of participation restriction; hence the distribution of the number of restrictions was skewed. The HADS (Zigmond & Snaith 1983), the SIP (Bergner et al 1981) and the SPPB (Guralnik et al 1994) all provide ordinal scores. Hypotheses 4 to 8 were therefore investigated graphically using box plots. Hypotheses 1 to 7 were tested in the NorStOP1 baseline dataset. Hypothesis 8 was tested in the CAS-HA dataset.

6.3.1.3 Criterion validity

There is no widely accepted 'gold standard' for the measurement of locomotor disability. Hence, the criterion validity of the locomotor disability-specific PF-10 items was not assessed.

6.3.1.3 Feasibility

<u>Respondent burden</u>

Respondent burden was assessed by reviewing previously published studies around the acceptability of the SF-36 as a whole. This review considering time taken to complete the questionnaire, the reading age required and the comprehensibility of the items to respondents. The methods used to find the studies containing these data are described below.

Box 6.2	Hypotheses regarding the construct validity of the locomotor disability-specific PF-10 items

1. In those people with self-reported,
   a. chest problems;
   b. heart problems;
   c. falls;
   d. breathlessness when walking;
   e. dizziness or unsteadiness;
   f. body mass index of 25 kg/m^2 or greater, i.e. overweight or obese;
   g. body mass index of less than 18.5 kg/m^2, i.e. underweight;

   5% more people will be "limited a lot" on each item than in those people without these problems, or with BMI between 18.5 and 24.9kg/m^2, i.e. normal weight.
2. 5% more people will be "limited a lot" in each item in those,
   a. reporting that their health is "fair" or "poor" than in those reporting their health to be " excellent", "very good" or "good";
   b. with any participation restriction, i.e. reporting restriction in one or more of 11 domains, than in those with no restriction;
   c. requiring aids or the assistance of others to move around inside the home than in those that do not require this assistance;
   d. requiring aids or the assistance of others go outside than in those that do not require this assistance;
   e. who live alone than in those who do not.
3. 5% more people will be "limited a lot" in each item in those who,
   a. go out for a walk;
   b. take a bath/shower;
   c. do heavy housework;
   d. do heavy gardening;
   e. do DIY;
   f. walk at least a quarter of a mile;
   g. walk at least two miles;

   less frequently than in those who do these activities more frequently.
4. In those reporting being "limited a lot" in each item, the average number of participation domains where a person is restricted will be highest. In those reporting being "not limited at all", the average number will be lowest.
5. In those reporting being "limited a lot" in each item, the average HADS anxiety score will be highest. In those reporting being "not limited at all", the average score will be lowest.
6. In those reporting being "limited a lot" in each item, the average HADS depression score will be highest. In those reporting being "not limited at all", the average score will be lowest.
7. In those reporting being "limited a lot" in each item, the average SIP alertness score will be highest. In those reporting being "not limited at all", the average score will be lowest.
8. In those reporting being "limited a lot" in each item, the average SPPB score will be lowest. In those reporting being "not limited at all", the average score will be highest.

Administrator burden

As with respondent burden, the burden of the locomotor disability-specific PF-10 items on the administrator was considered in terms of the administration of the whole SF-36 instrument and data were extracted from the literature in this field. When considering the individual locomotor disability-specific items, there is no need for a scoring process and so this aspect of administrator burden was not assessed.

Missing data

Missing data at the item level were considered in two ways in the NorStOP1. First, the average number and proportion of persons missing each item was calculated. Second, the average number and proportion of items missed per person in those who had at least one item missing was calculated. All analyses were conducted in the NorStOP1 baseline dataset, stratified by age and gender. This section also draws on the findings of the literature search described below.

6.3.1.4 Search of the literature

In order to obtain the studies for this review, a search of the literature was conducted based on citations of the original publication of the SF-36. For all relevant papers found, the reference lists were searched for further relevant studies.

6.3.2 Results

6.3.2.1 Agreement

Weighted Kappa statistics for all five locomotor disability-specific items from the PF-10 suggested that agreement between the test and retest responses was very good, according to the criteria of Altman (1991: pg 404) (Table 6.1). The relative responses to the five items on the test and re-test questionnaires are shown in Appendix C.

Table 6.1 Quadratic weighted Kappa values: agreement between individual PF-10 items in those who did not report a change in general health. NorStOP pilot study, test-retest component, n=131

Item	Kappa (95% confidence interval)
d (climb several flights of stairs)	0.85 (0.78, 0.91)
e (climb one flight of stairs)	0.87 (0.79, 0.93)
g (walk more than a mile)	0.81 (0.72, 0.87)
h (walk half a mile)	0.80 (0.70, 0.87)
i (walk 100 yards)	0.88 (0.80, 0.93)

6.3.2.2 Construct Validity

There were differences of at least 5% in the proportion of people reporting being "limited a lot" on all five locomotor disability-specific items of the PF-10 between those with and without self-reported chest problems, heart problems, a history of falls, breathlessness on walking and dizziness or unsteadiness. This magnitude of difference was also seen for those with low BMI (<18.5kg/m^2) compared to normal BMI for all five items and was true for items d (several flights of stairs), g (walk more than a mile) and h (walk half a mile) when considering higher (>25kg/m^2), compared to normal BMI.

In those with poor self-reported general health, any participation restriction, requiring aids or the assistance of others to move around inside the house or to go outside and those living alone, at least 5% more people reported that they were "limited a lot" in all five items than in those reporting good general health, those with no participation restriction, those not requiring aids or assistance or those not living alone. In all activities in Hypothesis 3, in those who reported doing the activity less often, at least 5% more people reported being "limited a lot" than in those who reported doing the activities more frequently. Detailed results of these hypothesis tests are given in Appendix C.

For Items e (climb one flight of stairs), h (walk half a mile) and i (walk 100 yards), there was a clear trend in the number of domains of participation restriction across item response categories. However, in Items d (climb several flights of stairs) and g (walk more than a mile) there was little difference between the number of restrictions in those "limited a little" and "not limited at all". There was a trend within each item, that at lower levels of limitation, the median HADS anxiety and depression scores and the SIP cognitive complaint scores were lower. For all items, the median SPPB score was lowest in the group reporting being "limited a lot" and increased in those "limited a

little" and "not at all limited", showing increasing levels of physical ability. Details results of the testing of Hypotheses 4 to 8 are shown in Appendix C.

Overall, 120 hypotheses were tested and 116 were upheld. A summary is shown in Table 6.2. According to the criteria of Terwee et al (2007), there is evidence in support of the construct validity of all five locomotor disability-specific items of the PF-10 as indicators of locomotor disability.

6.3.2.3 Feasibility

Respondent burden

Time taken to complete the tool has only been measured for the SF-36 as a whole. The developers suggest that it takes five to 10 minutes to complete in a general population (Ware 2000). However, this has been shown to be longer in older populations. In their sample aged 65 years and over, Andresen et al (1998) found a mean time for completion of 12.6 minutes (standard deviation 8.5 minutes). In a similarly aged sample of people attending general practice and outpatient clinics, Hayes et al (1995) reported completion times from four to 30 minutes. However, the median time to complete was eight minutes, with 84% completing in less than 10 minutes. McHorney (1996) provided age-stratified estimates of completion times showing an age-related increase in the length of time needed to complete the questionnaire (Table 6.3). However, in order for the time to have been recorded, the tool was usually completed in a clinic setting, rather than being sent by post. This may influence the completion times of the instrument reported in these studies.

Given that interest here is in five items from the PF-10, if each of the SF-36 items took equally long to complete (although it is not clear if this is the case), a time of 10 minutes to complete the whole tool would suggest around 1 minute and 25 seconds to complete these items. However, placement of items within the tool and the number of other tools included in the questionnaire booklet will affect the completion time, as respondents tend to slow down as they progress through the items (Andresen et al 1996).

Table 6.2 Summary of the results of testing the construct validity of the individual locomotor disability-specific PF-10 items in the NorStOP1 (n=7,878) and CAS-HA (n=623)

Hypothesis	Number of items in which hypothesis is upheld[a]
1. Self-reported	
a. chest problems	5
b. heart problems	5
c. falls	5
d. breathlessness when walking	5
e. dizziness or unsteadiness	5
f. body mass index >25 kg/m^2	3[b]
g. body mass index <18.5 kg/m^2	5
2. Self-reported	5
a. general health	5
b. any participation restriction	5
c. require aids or the assistance of others to move inside the house	
d. require aid or the assistance of others to go outside	5
e. live alone	5
3. Frequency of	
a. Go out for a walk	5
b. take a bath/shower	5
c. do heavy housework	5
d. do heavy gardening	5
e. do DIY	5
f. walk at least quarter of a mile	5
g. walk at least two miles	5
4. Number of participation domains	3[c]
5. HADS anxiety score	5
6. HADS depression score	5
7. SIP alertness score	5
8. SPPB score	5
Total n (%)	116 (96.7)

[a]Five analyses per hypothesis (one per locomotor disability-specific PF-10 item); [b]Evidence against Hypothesis for items e (one flight of stairs) and i (walk 100 yards); [c]Evidence against Hypothesis for Items d (climb several flights of stairs) and g (walk more than a mile)

Table 6.3 Time to complete SF-36, by age (McHorney 1996)

Age-group	Average time to complete questionnaire (minutes)
Less than 65 years	8.0
65 to 74 years	9.7
75 years and over	12.9

In terms of the comprehensibility of the SF-36 items, Ware et al (2000) suggested that the SF-36 should not be administered to those with limited reading ability, be this due to a lack of the English language or to illiteracy. Foreman and Kleinpell (1990, cited in McHorney 1996), recommended questionnaires should be written for a reading level of no higher than US grade five (age 10 to 11 years old) for surveys of elderly persons. However, McHorney (1996) states that the reading age of some of the SF-36 items is US grade seven (age 12 to 13 years old) or higher, and so up to a third of the elderly population might be expected to have difficulty with these items. It is not clear what reading age is necessary to understand the five items of interest here.

Other studies also provide evidence that older people may find completion of the SF-36 difficult. Mallinson (1998) found that 64% of her sample aged 65 years and over needed assistance from a friend or relative to complete the SF-36, whilst 28% said they had no problems completing it. In a follow-up qualitative study Mallinson (2002) found several potential explanations for the difficulty with completion. People with chronic illnesses, particularly of an episodic nature, did not know whether to answer in relation to their illness or not. Other problems included difficulty in determining what the distances of a mile, half a mile and 100 yards were referring to: they were seen as abstract and respondents failed to put them into the context of their daily lives. In contrast to the findings of Mallinson (1998, 2002), Hayes (1995) reported that 91% of participants found most or all of the SF-36 questions clear and easy to understand and Andresen et al (1998) reported that 66% of people were very or somewhat satisfied with the SF-36. The differences between these studies and those of Mallinson (1998, 2002) may lie in the purpose of the studies. The objective of Mallinson (1998, 2002) was to test the SF-36 from a qualitative perspective, whilst Hayes et al (1995) and Andresen et al (1998) considered clarity and satisfaction respectively in studies testing the quantitative properties of the instrument.

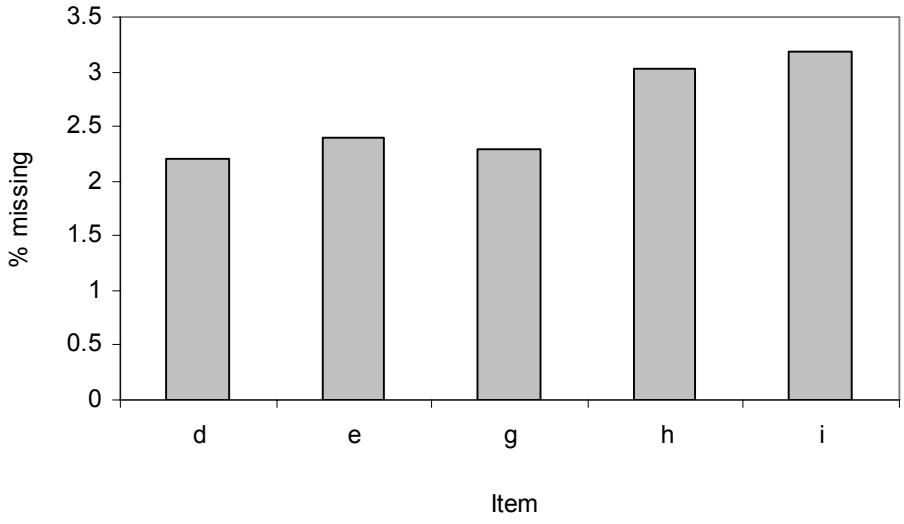
Administrator burden

The SF-36 is suitable for manual data entry or the use of scanning technology (Ware et al 2000), meaning that it is a flexible tool. However, different modes of administration and the inclusion of extra items could influence responses and so that results are not necessarily comparable (Bowling et al 1999).

Missing data

The percentage of people missing an individual item ranged from 2.2% (Item d - climb several flights of stairs) to 3.2% (Item i - walk 100 yards) (Figure 6.1). Levels of missing data were considerably higher (up to 8.1%) in both genders in those aged 80 years and over for Items h (walk half a mile) and i (walk 100 yards) than for younger responders (data not shown). These levels of missingness are broadly in line with those reported in the literature (McHorney et al 1994, Parker et al 1998).

Figure 6.1 Percentage of respondents missing each locomotor disability-specific PF-10 item in the NorStOP1 (%), n=7,878



d - climb several flights of stairs; e - climb one flight of stairs; g - walk more than a mile; h - walk half a mile; i - walk 100 yards

McHorney (1996) commented that PF-10 item responses are often non-randomly missing. This was reiterated by Gandek et al (1998), who noted the increasing level of item non-completion

found in some studies in those items that form hierarchical scales, i.e. the tasks get easier as the respondent moves through the items: climb several flights of stairs, climb one flight of stairs; walk more than a mile, walk half a mile, walk 100 yards. This was also found in the NorStOP1 (Figure 6.1), with higher levels of missingness in the 'easier' items in a hierarchy that appear later in the item set. This is likely to be because someone who can complete the first task within the hierarchy must logically be able to complete the remaining tasks and hence respondents feel that they do not need to complete the later items.

Of the 7,878 people in the NorStOP1, 7,492 (95%) completed all five locomotor disability-specific PF-10 items. The most common pattern of missing data was to have five items missing (105 respondents, 27%), although many people missed only one or two items. Levels of missing data were higher at older ages and in males. Table 6.4 shows the median number of missing items by age and gender in those who missed one or more items.

Overall, levels of missing data in the locomotor disability-specific items of the PF-10 were low, although there may be concerns over the patterns of missingness.

Table 6.4 Median (IQR) number of missing items in the NorStOP1 missing one or more the five locomotor disability-specific PF-10 items, n=386

	Males	Females	Overall
50 to 59 years	4 (1, 5)	2 (1, 5)	3 (1, 5)
60 to 69 years	2 (1, 5)	2 (1, 3.5)	2 (1, 5)
70 to 79 years	2 (1, 5)	2 (1, 4)	2 (1, 5)
80 years and over	2 (1, 5)	2 (1, 3.5)	2 (1, 4)
Overall	2 (1, 5)	2 (1, 4)	2 (1,5)

IQR=Interquartile range

6.3.3 Summary

In summary the individual locomotor disability-specific items of the PF-10 have been found to have good repeatability, assessed as agreement over time, and good construct validity. The PF-10 as a whole, and therefore the items in question are easy to administer, leading to a feasible set of items from the administrator point of view. Although reported not to be fully comprehensible to all potential respondents, the items have generally been reported to be acceptable to participants in quantitative studies. Levels of missing data are generally low, with the exception to this being in the

oldest age-groups, a group to whom the items may be less clear. For the majority of people, there is no evidence that the PF-10 items present a significant burden.

The major limitation of these five items, as noted above, is that they cannot measure locomotor disability, as they merely represent the level of limitation in a specific activity on three levels, i.e. they provide ordinal-level data. The next section considers the combination of items into a scale using the methodology employed to create the score for the full PF-10.

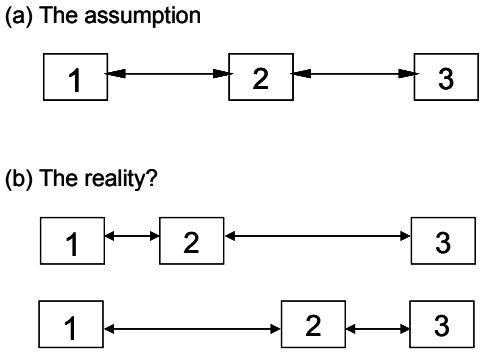
6.4 The use of summated ratings to create a measure of locomotor disability from the five items of the PF-10

In the original PF-10, the scale score is derived from the 10 individual items using an unweighted summated ratings method, i.e. allocate a score to the individual item responses and sum these scores to get an overall scale score. In the case of the PF-10, the raw score, which ranged from 10 to 30, was transformed to a 0 to 100 point scale using the formula $100\ \frac{\text{score}-10}{20}$. Under this scoring mechanism, 100 represents the best possible physical functioning and zero represents the worst possible functioning (Ware et al 2000).

In the case of the five items from the PF-10, each item is scored 1, 2 or 3 and so the resulting raw sum score would range from five to 15. Summated rating scales are unlikely to form interval-level measures that allow the calculation of a change in score over time or a difference in score between individuals or groups (Svensson 2001), and the empirical testing of this concept is problematic (Tennant 2007).

To illustrate this point, first consider a single item from the PF-10: walking 100 yards. The response options are “yes, limited a lot”, “yes, limited a little”, “no, not limited”. These options are scored 1, 2 and 3 respectively. In a summated rating approach, the responses are treated as though they were equally spaced, i.e. as if they were interval-level data (Figure 6.2(a)). However, this need not be the case and by the definition of ordinal data, we cannot know this to be the case (Merbitz et al 1989). In reality, the distance between adjacent categories is unlikely to be equidistant and hence could be represented by either of the examples in Figure 6.2(b).

Figure 6.2 Ordinal responses to an item under simple summated ratings



If the assumptions of summated ratings, i.e. Figure 6.2(a), are to be accepted, this directly implies that a change from a baseline score of 1 to a follow-up score of 2 is the same as a change from a baseline score of 2 to a follow-up score of 3 and that a change in score from 1 to 3 is the size of a change from 1 to 2 plus the size of a change from 2 to 3. However, if the summated rating assumption does not hold, i.e. Figure 6.2(b), then it cannot be assumed that the change in the underlying construct needed to move from a score of 1 to a score of 2 is the same as the change in score needed to move from a score of 2 to a score of 3.

This issue may seem trivial when considering a single item with a small number of response categories, for example any of the individual PF-10 items. However, it becomes more important when combining a series of ordinal-level scales to create a multi-item score. By summing multiple ordinal-level scales, an interval-level measure with the required mathematical properties is not achieved; instead a longer ordinal-level scale is generated. This scale, where scores of the same value may not imply the same response pattern for everyone with that score, may not be a useful representation (Merbitz et al 1989). In fact, Svensson (2001) suggests that even the summation of the ordinal scores from multiple items itself does not make sense.

Within the five locomotor disability-specific items, there is an additional objection to a summated ratings approach: logically a hierarchy exists within items. For example, a person should have a higher level of disability to report a particular degree of limitation in walking half a mile, than they should to report that level of limitation in walking more than a mile. This suggests that although all items may be equally important in terms of mapping to the locomotor disability construct, they cannot all be equally related to the underlying concept.

For these reasons, the summated ratings method will not be used to create a measure of locomotor disability. The next section does however present an initial investigation into whether the items might usefully be combined using another method.

6.5 The potential to create a measure of locomotor disability from the five items of the PF-10

Although the use of the summated ratings method to create measure of locomotor disability has been ruled out, the items themselves have been shown to have good measurement properties. This section considers the associations between the five locomotor disability-specific items of the PF-10 and hence their potential for combination into a measure. In particular the unidimensionality of the items and their response distribution, i.e. the empirical presence of a hierarchy, are evaluated.

6.5.1 Methods

6.5.1.1 Unidimensionality

Factor analysis

As discussed in Chapter 5, in order to be sensibly combined to produce a scale score, items should belong to a single dimension. The five locomotor disability-specific items of the PF-10 were assessed for their membership to a single dimension using confirmatory factor analysis (CFA) (for example, Byrne (2001)). This analysis specifies a correlation structure between the items, i.e. that they consist of a single dimension, and accounts for the ordinal nature of the individual item responses. CFA was carried out in the NorStOP baseline dataset using MPlus Demo Version 5.21 (Muthén & Muthén 2010). Five indices of model fit were considered: chi-square test, Comparative Fit Index (CFI), Tuker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) and Weighted Root Mean Square Residual (WRMR). The desirable ranges of these values are given in Table 6.5.

Internal consistency

Cronbach's alpha for the five locomotor disability-specific PF-10 items in the NorStOP1 was calculated using SPSS 16.0.

6.5.1.2 Item response distribution

As discussed above, there is a logical hierarchy within the items. This hierarchy was assessed empirically in two ways. First, item responses were displayed graphically. Second, relative responses to individual pairs of items within each hierarchy were considered in cross-tabulations. These analyses were carried out in the NorStOP1 baseline dataset.

6.5.2 Results

6.5.2.1 Unidimensionality

Factor analysis

There is evidence against the fit of the five locomotor disability-specific items of the PF-10 to a single factor from the chi-square test, RMSEA and WRMR calculated from the confirmatory factor analysis (Table 6.5). The fit is acceptable according to the CFI and TLI. Although the poor fit based on the chi-square test could be ascribed to the large sample size, there is still little statistical evidence in favour of the hypothesis that the five items belong to a single dimension.

Table 6.5 Confirmatory factor analysis of the five locomotor disability-specific items from the PF-10 in the NorStOP1, n=7,492

Fit statistic	Desirable range	Value in NorStOP1
Chi-square test of model fit[a]	p>0.05	<0.0001
CFI	>0.9	0.996
TLI	>0.9	0.994
RMSEA	<0.05	0.109
WRMR	<1.0	3.026

[a]Highly influenced by sample size; CFI - Comparative Fit Index; TLI - Tuker-Lewis Index; RMSEA - Root Mean Square Error of Approximation; WRMR - Weighted Root Mean Square Residual

Internal consistency

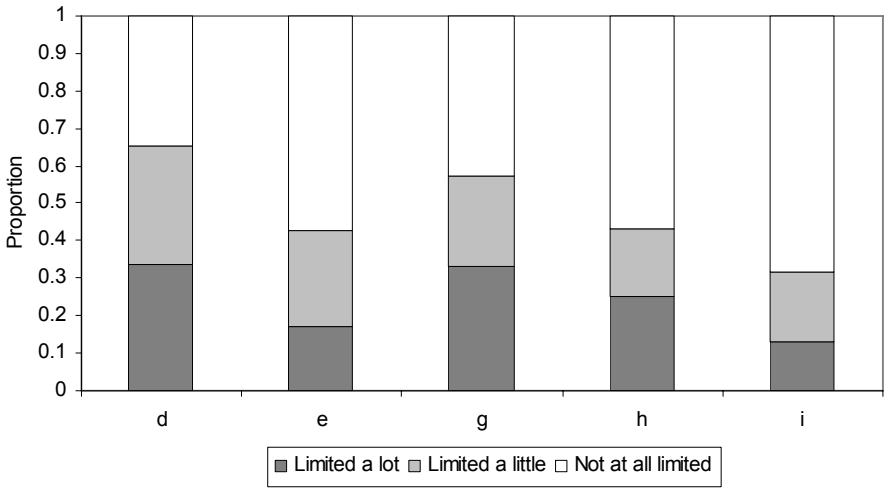
Cronbach's alpha for the five locomotor disability-specific PF-10 items in the NorStOP1 was 0.938, indicating possible item redundancy. Given the results of the CFA, this high value of alpha could also indicate the presence of two correlated subscales, for example related to walking and stair-climbing.

6.5.2.2 Item response distribution

Figure 6.3 shows that people report higher levels of limitation in climbing several flights of stairs than in climbing one flight and in walking longer rather than shorter distances. This confirms the proposed hierarchy of the items in the NorStOP1.

Consideration of the relative responses to items within the proposed stair-climbing and walking hierarchies (Appendix C) revealed that in the case of each proposed hierarchy, less than 1% of respondents reported a higher level of limitation in the item involving a shorter rather than longer stair-climb or walk. Again, this lends empirical evidence to the idea of a hierarchy within the items.

Figure 6.3 Locomotor disability-specific PF-10 item response distributions in the NorStOP1, n=7,878



d - climb several flights of stairs; e - climb one flight of stairs; g - walk more than a mile; h - walk half a mile; i - walk 100 yards

6.5.3 Summary

There is evidence in the form of Cronbach's alpha that the five items are all assessing the same construct. However, this statistic can be inflated by the presence of two subscales within the items. This is a possibility, given the results of the CFA, which did not confirm the presence of a single dimension. This high value of alpha could also suggest item redundancy, with two or more of the five items assessing the same construct with evidence of a hierarchy to the item responses within the walking and stair-climbing item groups, this seems likely.

6.6 Discussion

6.6.1 Principal findings

This chapter has investigated the potential usefulness of the PF-10 as a measure of locomotor disability, as suggested by Syddall et al (2009), and previously used by Peat et al (2006a). Initial consideration of the PF-10 as a whole led to the conclusion that the items in the original instrument were too wide-ranging, with some items being a poor fit to the definition of locomotor disability adopted in this thesis (Section 6.2). For this reason, five of the items were excluded from further consideration and the five remaining items, those specific to locomotor disability (d - climb several flights of stairs, e - climb one flight of stairs, g - walk more than a mile, h - walk half a mile, i - walk 100 yards), were considered individually as markers of this construct (Section 6.3).

The five items cover the two main locomotor tasks included in most studies of locomotor disability identified in Chapter 3, within the definition of locomotor disability set out in Chapter 1. Although these items do not cover some of the less commonly included locomotor disability tasks, they do include the core tasks used in the majority of studies, giving confidence that the items can give an overview of locomotor disability in respondents to these items.

These five items were shown to have good repeatability. Agreement of item responses over four weeks in those who did not report a change in general health was very good, according to the criteria of Altman (1991: pg 404). The face validity of the individual items was considered to be good, as many previous studies have asked about walking and stair-climbing in relation to locomotor disability (see Chapter 3). There was also evidence of the construct validity of the items

from the testing of pre-specified hypotheses, 97% of which were upheld. Levels of missing data at the item level were low, although there was not an even spread of missingness, with older people and males more likely to miss items. There was also an uneven spread of missingness across items, with 'easier' items within the proposed walking and stair-climbing hierarchies more likely to be missed. A review of previous studies on the feasibility of the SF-36 showed that there is conflicting evidence surrounding the ease of completion of the locomotor disability-specific items from the PF-10. Some studies suggested that some items may be confusing or irrelevant to older respondents, whilst others reported the items to be acceptable.

Having established the individual items as good indicators of locomotor disability, Section 6.4 considered the combination of these items into a measure of locomotor disability using the unweighted summated rating method of the original PF-10. However, there was some evidence to suggest that the five locomotor disability-specific items may not form a single dimension, which is regarded as a necessary condition for the formation of a measure. Pursuit of the unweighted summated rating method of item combination was considered to be unsuitable, due to the proposed hierarchies within the items and theoretical production an ordinal-level scale, rather than an interval-level measure. The items were then tested for their fit to a single dimension. Results of the confirmatory factor analysis provided evidence against a single dimension and assessment of items responses suggested the presence of hierarchies within the walking and stair-climbing item groups.

6.6.2 Strengths and weaknesses of the study

The literature around the PF-10 mainly originated from studies involving the SF-36 as a whole, and so it was sometimes difficult to interpret the findings of these papers in terms of their relevance to this thesis. However, much of the qualitative work around the SF-36 concentrated on individual items and the understanding of these items by respondents, for example Mallinson (1998). This made interpretation of the results of these studies easier in terms of individual items than the more quantitative studies, for example Hayes et al (1995) and Andresen et al (1998), which considered the acceptability of the instrument as a whole. This difference in the ease of interpretation could have led to a bias in the consideration of the value and conclusions of these studies. As the qualitative studies of Mallinson (1998, 2002) were more negative about the acceptability of the SF-36 to respondents, than were the quantitative studies. It may be that the

interpretation of the acceptability presented in this chapter is biased towards concluding that the comprehensibility of the items is poor.

Considering the new analyses of the individual locomotor disability-specific items in this chapter, Kappa statistics were used to assess repeatability of items over time. Whilst these statistics are generally accepted to be the most appropriate for this type of analysis (Sim & Wright 2005, Terwee et al 2007), it is also widely acknowledged that Kappa statistics are strongly influenced by the prevalence of the attribute in question. With two response options, when prevalence close to 50%, the values of Kappa tend to be higher than when prevalence is closer to 0% or 100% (Sim & Wright 2005). However, as the purpose of Kappa is to correct for agreement by chance, and as chance agreement is higher when prevalence is higher (or lower) than 50%, it is inevitable that Kappa will be lower. This can be considered a result of the aim of the statistic and not necessarily a limitation (Vach 2005). Another criticism of Kappa is the arbitrary nature of the defined cut-off values. Although well recognised, these values have no tangible meaning as Kappa is not measured in the units of the scale. Finally, in this chapter, Kappa was calculated using quadratic weighting (for example, Sim & Wright 2005). This method was chosen because it is in common usage and it makes sense to give higher weight to 'closer' values. Should a different weighting scheme, or indeed no weighting have been applied, the results would have changed slightly. For example linear weights would have reduced the Kappa values somewhat, although not to the extent that conclusions regarding the repeatability of the item responses would have altered.

The testing of construct validity conducted in this chapter was undertaken using the methodology of Terwee et al (2007), whereby hypotheses were formed and tested in the NorStOP1 dataset. A high percentage of hypotheses were upheld, suggesting that these items are measuring what they are hypothesised to measure, i.e. locomotor disability. However, it is possible, although unlikely, that both the individual items and the measures they are tested against, for example frequency of activities, are assessing the same construct, but that this construct is not locomotor disability. The cut-off of 75% of hypotheses being upheld, as suggested by Terwee et al (2007) is arbitrary and these authors give no guidance as to the number of hypotheses it is necessary to test, but in this case, a much higher percentage than suggested were upheld. This would indicate that the items can be seen as valid assessments of locomotor disability. Terwee et al (2007) also suggest that hypotheses should be precise, specifying expected differences between groups, or the magnitude of expected correlations. In this study, an arbitrary difference of 5% was chosen as

the cut-off to define differences between groups, as there was no preliminary evidence on which to base an expected difference a priori.

Low levels of missing data were found in responses to the PF-10 items, but it should be noted that the missing data rates are within the responders to the NorStOP1 survey. Hence some of those likely to have high levels of missing data may have selected themselves out of the denominator by not responding to the questionnaire. This is likely to occur in any population survey and it is not possible to differentiate those who did not return the questionnaire specifically due to difficulties with the items of interest from those who did not return it for other reasons, such as difficulties with other scales in the questionnaire, or its overall length. It seems unlikely that in as large a questionnaire as those used in the NorStOP, the five items from the PF-10 would have substantially influenced overall response rates.

A further consideration in the new analyses presented in this chapter should be the generalisability of the findings to populations outside the NorStOP. Chapter 4 showed that the NorStOP is broadly representative of the population of England, although there is some evidence that males, those who live alone and those in lower socioeconomic groups might be underrepresented in the cohort. This sort of response bias is almost inevitable in studies of this type and will only be problematic if there is reason to suspect that those people who chose not to take part in the NorStOP may complete the questionnaire differently to those who take part.

6.6.3 Strengths and weaknesses of the study in relation to the current literature

The decision to reduce the PF-10 to five items for the purpose of the assessment of locomotor disability was taken on a theoretical basis: some of the items in the original PF-10 did not relate conceptually to the construct of locomotor disability as defined in this thesis. This decision is not supported by a previous factor analysis of the SF-36 that suggested that the ten items formed a single dimension (Garratt et al 1993). However, previous analyses employed exploratory factor analysis to find the best factor structure in the dataset being analysed. As the factor structure of the SF-36 was already hypothesised, it could be argued that a confirmatory factor analysis, to test the plausibility of this structure would have been more appropriate. Furthermore, these previous analyses required the assumption that individual item responses were on an interval-level scale. This is not the case, with each item response being on a three-level ordinal scale, and such an assumption may lead to erroneous conclusions (Merbitz et al 1989). In

this chapter, the five locomotor disability-specific items were hypothesised to form a single dimension and this was tested using a confirmatory factor analysis that allowed for the ordinal-level of the items responses. This analysis did not provide strong evidence in favour of the five items forming a single dimension and further calls into question the results of analyses of the whole SF-36 using exploratory factor analysis assuming interval-level item responses. It does however, gain some support for previous analyses of the PF-10 using Rasch analysis (Rasch 1960), which suggested a lack of unidimensionality in the 10 items when properly accounting for the ordinal-level nature of the item responses (Haley et al 1994, Jenkinson et al 2001).

The responsiveness (Bombardier & Tugwell 1987, Terwee et al 2007) of the individual items was not assessed, because the three-point ordinal-level scale of each item response is likely to be too coarse to provide a responsive instrument. Furthermore, the ordinal nature of the item responses limits the potential mathematical operations that are possible to investigate change. Assessment of responsiveness would be easier and provide more relevant information if measurement of locomotor disability were on an interval-level. A similar argument applies to the derivation of a MIC or a MCID, in that it does not make sense in the context of an individual item, either in terms of the coarseness of the measure or the ordinal level of this measurement. Furthermore, the determination of a MIC or MCID requires a benchmark of clinically relevant change against which this could be assessed, for example a global change question related to locomotor disability. An appropriate evaluation of global change was not available in the NorStOP and so this property could not be assessed.

It is also usual to present floor and ceiling effects (Terwee et al 2007), i.e. the proportion of the sample that obtained the highest or lowest score on an instrument. Again, this does not appear sensible when considering only a three-point ordinal response, as by the nature of there being only three categories, it is inevitable that large numbers of people will respond at the ends of the scale, i.e. report being “limited a lot” or “not limited at all”.

The final notable omission from this chapter is a reference to the interpretation of responses to the five items (Terwee et al 2007). Again, this is due to the nature of the individual item responses. The distribution of responses to the individual items in the NorStOP pilot study is given in Appendix C. However, further information on interpretation such as mean scores for an item would be inappropriate due to the mathematical operations required to calculate such statistics.

6.6.4 Meaning of the study

This chapter has shown the PF-10 to be a flexible instrument from the administrative point of view, both in terms of data collection and data entry, which can be carried out in a number of ways. The SF-36 has also been formally translated and adapted for use in different cultural settings, which will be of use in international studies. In addition, missing data levels are reasonable. These factors combine to make the PF-10 an attractive option for the measurement of locomotor disability and it has previously been suggested as a good measure of this construct (Peat et al 2006a, Syddall et al 2009).

There is currently a plethora of health assessment instruments available in the published literature and indeed Feinstein et al (1986) and de Vet et al (2003) suggest that there are too many. It is therefore desirable, where possible, to use or adapt existing instruments, rather than develop new sets of items. The SF-36 is the most widely used quality of life assessment instrument in the world (Garratt et al 2002), and so the development of items from this instrument into a measure of locomotor disability would be preferable to the development of a new set of items. This chapter has shown that five items from the PF-10 represent good individual indicators of locomotor disability from which responses can be reproduced over time. Furthermore, when considered in combination, these items present good coverage of the core activities related to the concept of locomotor disability. However, because the scoring mechanism suggested for the PF-10 does not produce a measure of locomotor disability, as defined in Chapter 2, these items have yet to be combined in an acceptable manner and empirical testing of the unidimensionality of the items suggests this may be unwise. It remains to be seen however, whether the similarity of items within a hierarchy may be causing the apparent multidimensionality: items within the hierarchy may be so similar to each other as to appear to form a different construct from the items in the other hierarchy.

6.6.5 Unanswered questions and future research

Two major questions remain following the analyses presented in this chapter, i) how useful are single items from the PF-10 in assessing locomotor disability in an epidemiological setting, and ii) how might these individual items be combined into a measure of locomotor disability?

First, the individual locomotor disability-specific items of the PF-10 have been shown to be repeatable, valid and feasible indicators of locomotor disability, but how useful might they be in assessing locomotor disability in an epidemiological setting? The walking items from the PF-10

have previously been used by Bohannon et al (2004) in a comparison of self-reported disability with measured performance. To build on this work in the setting of a postal survey, Chapter 8 considers the PF-10 items as indicators of locomotor disability in a cross-sectional analysis of the association between reports of pain and the level of disability. It seems likely however that this cross-sectional use of the items as markers of locomotor disability will be the extent of their usefulness in the context of epidemiological studies of locomotor disability. As discussed in this chapter, they provide only coarse levels of disability and do not provide for the detailed level of measurement required to assess changes over time.

Second, it would be useful to combine the items into a measure of locomotor disability. This would have several advantages. In the first instance, the combination of items would allow a finer grading of disability. Should this combination result in a true measure of locomotor disability it would allow the examination of change in the level of locomotor disability between groups and over time, and also the description of the distribution of disability in a population. In addition, it would mean that more of the construct of locomotor disability is covered by a single summary value, i.e. a person's score on the measure. This would be advantageous in the sense of the content validity of the measure being wider than that of an individual item, but also make administration of such analyses easier, as one would not have to repeat analyses for different outcomes, or choose between the different items for use as an outcome. Chapter 10 presents the use of Rasch analysis (Rasch 1960), the only proven way to create an interval-level measure from a set of ordinal item responses (Fischer 1995), to combine the five locomotor disability-specific items and simultaneously test the appropriateness of this combination, particularly in terms of the dimensionality of the items.

A further and more minor point in terms of work leading from this chapter is to note that all analyses presented here were conducted in the NorStOP cohort and so, although results should be broadly generalisable to the English population, further work is needed before they can be generalised to international or patient populations. Given the number of studies worldwide using the SF-36, this may well be worthwhile.

6.6 Conclusions

This chapter has described the potential use of the PF-10 as a measure of locomotor disability. It was found unsuitable for this purpose in two respects, i) some items were outside the

definition of locomotor disability adopted for this thesis, and ii) the scoring mechanism employed in the PF-10 does not create an interval-level measure. The individual items however were found to be good indicators of locomotor disability. The next chapter considers these items in a cross-sectional analysis of the association between socio-demographic factors, pain and locomotor disability using ordinal regression techniques suitable for the ordinal-level nature of the item responses. Chapter 10 presents the combination of the items into a measure of locomotor disability using Rasch analysis.

7 Approaches to modelling ordinal outcome data

7.1 Introduction

The previous chapter identified five items from the PF-10 as being good indicators of locomotor disability. Whilst these items do not constitute the interval-level measure argued for in earlier chapters, their ordinal item responses do present a more fine-grained approach to the assessment of locomotor disability than presented by the many dichotomous definitions of disability used in previous studies (see Chapter 3).

In order to assess the potential usefulness of this more fine-grained approach to assessment, the next chapter uses these ordinal responses to the individual PF-10 items as indicators of level of locomotor disability and models their association with socio-demographic factors and pain cross-sectionally in the NorStOP baseline cohort.

Ordinal outcomes present certain statistical challenges when they are to be included in a regression model. This chapter therefore considers the range of possible regression modelling techniques for these ordinal item responses. The advantages and disadvantages of each are considered along with their suitability in different scenarios.

7.2 Possible regression modelling approaches

There are many possible approaches to modelling ordinal data as the outcome in a regression model. Some are specific to ordinal data whilst others are not, and so some are more appropriate than others. This section describes the possible approaches and the suitable uses for each modelling approach.

7.2.1 Approaches not specific to ordinal data

This section describes approaches to modelling ordinal data that are not specifically designed for this purpose, but nevertheless are often employed with ordinal outcomes.

7.2.1.1 Ordinary least squares regression

One way of analysing an ordinal categorical variable, such as responses to the PF-10 items, is to treat it as though it is an interval-level variable. This assumption is almost certain to be violated with ordinal data (Armstrong & Sloan 1989). However, this is not the only consideration: modelling techniques for continuous data, such as an ordinary least squares regression, make the assumption of homogeneous variance. This is unlikely in an ordered categorical response (Anderson 1984), especially when the variable consists of qualitative categories (see Chapter 2), rather than being grouped continuous. Thus employing techniques for interval-level data to ordinal-level data can result in incorrect inferences and conclusions (Scott et al 1997).

7.2.1.2 A single dichotomisation

An alternative to treating the dependent variable as though it is continuous is to dichotomise it. This is an appealing idea, and examples of this can be found in the literature (Avlund et al 2003, Sainio et al 2006). The binary logistic regression model can be fitted in most standard statistical packages and the interpretation of the estimated odds ratios is well understood. Unlike treating ordinal-level data as though they are interval-level, dichotomisation is a valid approach. Manor et al (2000) found that with a large sample size, similar results arose from a binary logistic regression model, as from a model formulated specifically for ordinal data. However, these authors and others acknowledge that dichotomisation does leads to a loss of information and statistical power (Ananth & Kleinbaum 1997, Manor et al 2000). Furthermore, Scott et al (1997) discussed how arbitrarily choosing a dichotomy ignores relationships between odds ratios produced at different possible dichotomisations, and can lead to effect estimates that are relevant only at that point of dichotomisation. This is not to say though that the dichotomy should be chosen having first looked at the data, as this can lead to bias in the effect estimates (Campbell 2001: pg 89).

7.2.1.3 A sliding dichotomy

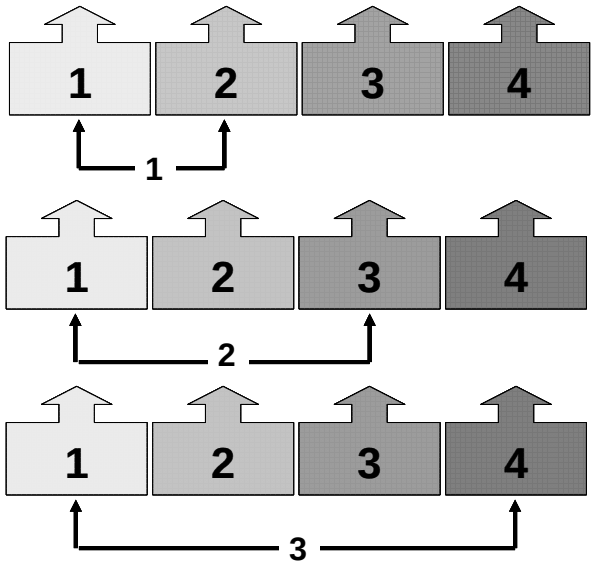
Extending the idea of dichotomising the data at a single point is the analysis of the data at all possible dichotomies, using a series of binary logistic regression models. Again, this approach is valid and can be carried out in standard statistical packages. It also makes more use of the data than a single dichotomisation. However, this is not a particularly efficient method of analysis and

large numbers of parameters are produced, particularly when there are several response categories and/or independent variables. This can make the reporting of results cumbersome and as a consequence, difficult to interpret, especially when there is little difference between the estimates of the odds ratios at each dichotomisation.

7.2.1.4 The multinomial logistic regression model

The multinomial logistic regression model (MLRM), also known as the polytomous or polychotomous logistic regression model or the discrete choice model (Hosmer & Lemeshow 2000), is able to deal with more than two response categories. In the MLRM, each category is compared to a reference category, as in Figure 7.1, through the use of logits to estimate odds ratios (see Box 7.1). Due to the nature of the comparison that it makes, this model ignores any ordinality in the data. Whilst it is not incorrect to use the MLRM with ordinal-level data, it does not take full advantage of the ordinal-level nature of the outcome variable, and so some information regarding the association between the independent and dependent variables is lost, reducing statistical power.

Figure 7.1 Illustration of the comparisons made in a multinomial logistic regression model: an example with four response categories



Box 7.1 Mathematical description of the multinomial logistic regression model

Let Y be a categorical response variable with k unordered categories and x_i the independent variable x at level i. Let $p_{ij} = P(Y = j \mid x_i)$

The model equations are,

$$\log_e \frac{P(Y = j \mid x)}{1 - P(Y = 1 \mid x)} = \alpha_j + \beta_j x \text{ , } j=1,\ldots,k.$$

Given the logistic form of the model, the p_{ij} are calculated as

$$p_{i1} = \frac{e^{x_i\beta_1}}{e^{x_i\beta_1} + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}, \quad \ldots, \quad p_{ik} = \frac{e^{x_i\beta_k}}{e^{x_i\beta_1} + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}$$

However, these equations do not identify the model (give a unique solution) because there are k parameters to estimate from k equations. To make the model identifiable, one of the ß-parameters, $ß_j$, is constrained to be equal to zero. Choosing the first category as the reference group and hence constraining $ß_1$=0, the set of equations becomes

$$p_{i1} = \frac{1}{1 + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}, \; p_{ik} = \frac{e^{x_i\beta_k}}{1 + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}, \quad \ldots. \quad , \quad p_{ik} = \frac{e^{x_i\beta_k}}{1 + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}$$

Then, the probability of outcome $J=j$ relative to the reference category where J=1 is,

$$\frac{p_{ij}}{p_{i1}} = \frac{\frac{e^{x_i\beta_j}}{1 + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}}{\frac{1}{1 + e^{x_i\beta_2} + \ldots + e^{x_i\beta_k}}} = e^{x_i\beta_j}$$

Hence, $logit(P(Y=j)|\underline{x})$ can be interpreted as the odds of being in outcome category j compared to outcome category 1, for a one unit increase in the independent variable x.

In the case of p independent variables, this becomes,

$$\log_e \frac{P(Y = j \mid \underline{x})}{1 - P(Y = 1 \mid \underline{x})} = \alpha_j + \beta_{1j} x_1 + \beta_{2j} x_2 + \ldots + \beta_{pj} x_p \text{ , } j=1,\ldots,k.$$

7.2.2 Ordinal modelling approaches

Aside of the non-ordinal modelling approaches described in the previous section, there are techniques that were derived for the analysis of ordinal-level data. All of the methods discussed in this section are extensions of the binary logistic regression model.

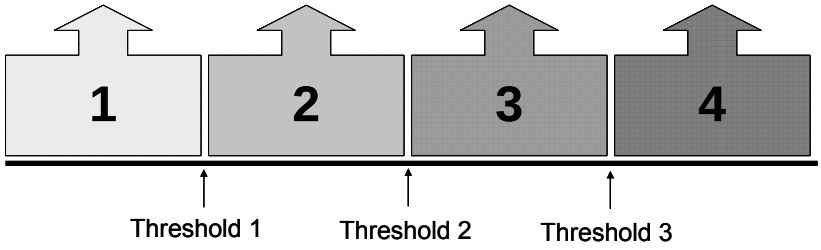
7.2.2.1 The generalised ordered logit model

This model is very similar to the use of the sliding dichotomy described above, and can be interpreted in the same way. The difference between the two approaches is that with this model, all of the parameters, for each of the dichotomies, are estimated simultaneously.

7.2.2.2 The proportional odds model

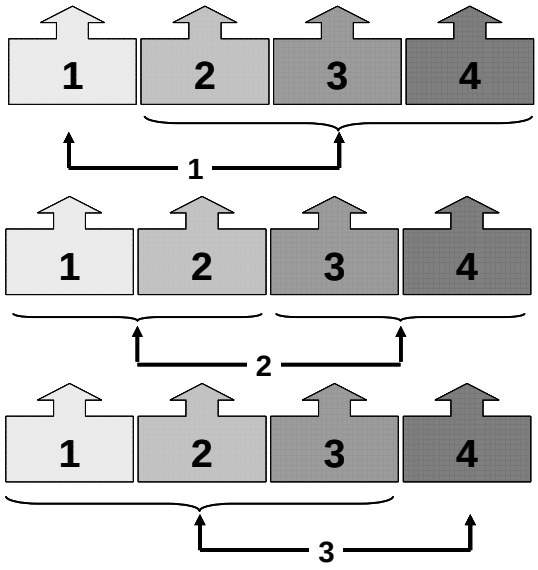
The proportional odds model (POM), also known as the cumulative odds model (Manor et al 2000) or the cumulative logit model (Ananth & Kleinbaum 1997), is the most popular ordinal logistic regression model (Bender & Grouven 1997, Lall et al 2002). The development of the POM was motivated by the idea of modelling grouped continuous data and initially is best considered in this context. Figure 7.2 illustrates the idea of a continuum and how it is partitioned by thresholds.

Figure 7.2 An illustration of the continuum of a dependent variable that can be partitioned by thresholds: an example with four response categories



The POM is based on the idea of modelling cumulative probabilities (Figure 7.3), and is not dissimilar to the idea of a sliding dichotomy or generalised ordered logit model. However, the POM fits the series of binary logistic regression models simultaneously and assumes that the estimate of the β-coefficient, and hence the odds ratio, for each independent variable in each regression equation is the same (Box 7.2). This is known as the proportional odds or parallel lines assumption and implies that the effect of the independent variable is the same at each level of the dependent variable.

Figure 7.3 Illustration of the comparisons made in a proportional odds model: an example with four response categories



As the POM was motivated by the idea of grouped continuous data (Greenland 1994, Campbell 2001: pg 90, Lunt 2001a), it is considered to be particularly valid when the dependent variable is directly related to a latent variable (Bender & Grouven 1997), i.e. some underlying continuous phenomenon. Indeed, some authors (Anderson 1984, Lall et al 2002) consider that, as the α-parameters (see Box 7.2) represent the thresholds on the latent scale, the model becomes very difficult to interpret when the outcome is not grouped continuous. However, others concede that the model can be useful in circumstances when the outcome is a series of qualitative categories (Armstrong and Sloan 1989, Stromberg 1996). Indeed, it has been suggested (Long & Freese 2006: pg 187) that the POM, as presented here, can be viewed as a non-linear probability model, thus a latent variable is unnecessary for interpretation, and the POM can be used with all forms of ordinal-level data.

This model has two main advantages over other models for ordinal data. First it is easy to fit in most standard statistical packages and it is easy to present and interpret the results (Bender & Grouven 1997), as no more odds ratios are estimated than for a binary logistic regression model.

Box 7.2 Mathematical description of the proportional odds model

Let Y be an ordinal categorical dependent variable with k categories. There are then k-1 thresholds between categories. Let x_i the independent variable x at level i.

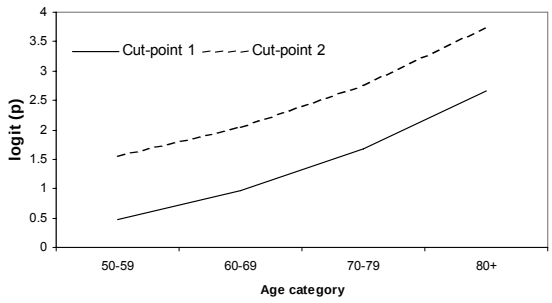
Let $p_{ij} = P(Y = j \mid x_i)$ and $\pi_{ij} = P(Y \leqslant j \mid x_i)$. Then $p_{i1} + p_{i2} + \ldots + p_{ik} = \pi_{ik} = 1$

A binary logistic regression model uses a logit link function to model the probability that an individual is in each category. Then dichotomising the data at each threshold and fitting k-1 binary logistic regression models gives,

$$\log_e \frac{\pi_{i1}}{1-\pi_{i1}} = \alpha_1 + \beta_1 x_i, \quad \ldots, \quad \log_e \frac{\pi_{ik\text{-}1}}{1-\pi_{ik\text{-}1}} = \alpha_{k\text{-}1} + \beta_{k\ 1} x_i, \ i=1,\ldots,p, \ j=1,\ldots,k\text{-}1$$

The POM fits these equations simultaneously and assumes that $ß_1 = ß_2 = \ldots = ß_k$. This implies that the logits estimated by the model are parallel, and the assumption is often called the parallel lines or proportional odds assumption.

An example of the parallel lines assumption with age-group as the independent variable



Hence, for a single independent variable, x, the POM is modelled by the equation

$$\log_e \frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_j + \beta x_i, \ j=1,\ldots, k\text{-}1,$$

where the ß-parameter does not depend on j, the level of the dependent variable.

In the case of p independent variables, this becomes

$$\log_e \frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_j + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \ j=1,\ldots, k\text{-}1,$$

where each ß-parameter is independent of the level j, of Y.

Second, the model is invariant to, i.e. is not changed by, the reversal of category order (Armstrong & Sloan 1989, Ananth & Kleinbaum 1997, Campbell 2001, Lunt 2001a) or by the collapsing of adjacent categories (Greenland 1994, Manor et al 2000). The reversal of the ordering of the categories results only in a change in the sign of the model coefficients (Ananth & Kleinbaum 1997). This can be particularly useful when there is no firm belief about which direction it makes most scientific sense to model (Armstrong & Sloan 1989). For example, is it best to model the probability of better health or of worse health? Collapsibility invariance means that if adjacent categories are merged, the β-coefficient for each independent variable will not change. However, the estimate of the β-coefficient is affected by the size of the sample, deviations from the proportional odds assumption and the placement of the thresholds on the continuum (Stromberg 1996) and so in practice, merging adjacent response categories will result in slightly different estimates of β. Some authors do not see invariance as a positive property of the POM. Greenland (1994) critically discussed how it may not be beneficial to assume invariance and collapse categories if the underlying data generating process is not invariant. Manor et al (2000) regarded the collapsibility assumption as implying that a latent variable is being modelled, hence strengthening the case that this model can only be used with grouped continuous data.

The major disadvantage of the POM is the strong assumption that the effect of the independent variables is the same at each level of the dependent variable. If this assumption is violated, model fit can be poor and inferences misleading (Lunt 2005). For this reason, many authors do not consider the POM to be useful in practice. For example, Long and Freese (2006; pg 200) observed that "*the parallel regression assumption is frequently violated*". When the assumption is violated, it is clear that the POM should not be used (Lunt 2005, Long & Freese 2006). Bender and Grouven (1998) demonstrate, using an example of the effect of smoking on diabetic retinopathy, that using the POM where the assumption is violated can lead to incorrect conclusions.

However, it is debatable how best to test for the assumption of proportional odds and how to decide when it has been met. A chi-square score test can be used to test for deviations from the model assumptions, however this tends to be anti-conservative, i.e. gives a more significant result than is true (Ananth & Kleinbaum 1997). This is especially true for moderate departures from the assumption (Lall et al 2002). Some packages, for example Stata (StataCorp LP 2009), offer a global likelihood ratio test, comparing the POM to a series of binary logistic regression models

(Long & Freese 2006: pg 199). Another indication of the variables for which the assumption of proportional odds fails is a Wald test (Long & Freese 2006: pg 199). Alternatively, nested models can be compared using likelihood ratio tests, to assess the proportionality of individual or groups of variables. As with most statistical tests, those assessing the proportionality of odds are affected by sample size and so in a large sample, a rejection of the assumption by a formal statistical test is more likely. Campbell (2001: pg 92) proposed that the POM is robust to mild departures from the proportional odds assumption and Cole (1999) suggested that an alternative to the POM is only required when there is a qualitative break in the estimated regression coefficients, i.e. the direction of the effect changes, meaning that estimates of β-coefficients change from positive to negative or vice versa. Overall, it seems that the main assumption of the POM is often violated and that in practice, alternative, more flexible models are required.

7.2.2.3 The partial proportional odds model

The partial proportional odds model (PPOM) is similar to the POM, but it allows the assumption of proportional odds to be relaxed for q of the p independent variables (Box 7.3) (Ananth & Kleinbaum 1997, Lall et al 2002). Williams (2006) suggested that the PPOM can be viewed as a compromise between the parsimony, but strong assumptions, of the POM and the large number of parameters in the MLRM. If $p=q$ then the model has the same number of parameters as the MLRM, hence the advantage of parsimony in the PPOM is lost and the model is equivalent to the generalised ordered logit model.

The general version of the model is called the 'unconstrained' PPOM. This means that there are no constraints placed on how the odds ratios vary over thresholds (Lall et al 2002). A 'constrained' PPOM uses a set of constraints to impose a linear relationship on the odds ratios for variables in the model that do not meet the proportional odds assumption (Ananth & Kleinbaum 1997, Lall et al 2002). This set of constraints is the same for each independent variable (Lall et al 2002). Constraints can be difficult to derive and are only necessary if there are a priori assumptions about how the data should behave. In this thesis the term PPOM is used to refer to the unconstrained model.

The main advantage of the PPOM over the POM is the flexibility to allow different effects of independent variables at different levels of the dependent variable (Lunt 2005). This is a property that Anderson (1984) referred to as "*multidimensionality*". However, because the PPOM allows the

assumption of proportional odds to be violated for some variables, the non-parallelism of the logits means that it is possible to predict negative probabilities (Lunt 2001b). This may not be a problem in studies where the aim is to investigate associations between variables, but it would be problematic if the aim of the analysis was to generate a predictive model. In many cases, negative probabilities will only be predicted outside of the observed data range and, as in general, it is not regarded as good practice to predict outside the range of the original dataset, this should not be a concern (Williams 2006). Williams (2006) also suggested that negative predicted probabilities are usually associated with other problems with the model, for example high standard errors or overly complicated models.

Box 7.3 Mathematical description of the partial proportional odds model

As in Box 7.2, assume that Y is an ordinal categorical dependent variable with k categories. There are then k-1 thresholds between categories.

Let there be p independent variables x_i, and then the equations for the POM are,

$$\log_e \frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_j + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \quad j=1,\ldots,k-1$$

In the POM, the assumption is that the slope parameter, ß, is constant across all outcome categories for a particular independent variable.

The PPOM allows the assumption to be relaxed for q of the p independent variables. Hence, when p=3 and x_1 and x_2 show proportionality whilst x_3 fails the assumption, the equations are

$$\log_e \frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_j + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{3j} x_{3i}, \; j=1,2,\ldots,k-1$$

For x_3, a separate ß-parameter, $ß_{3j}$, is estimated for each of the k-1 thresholds.

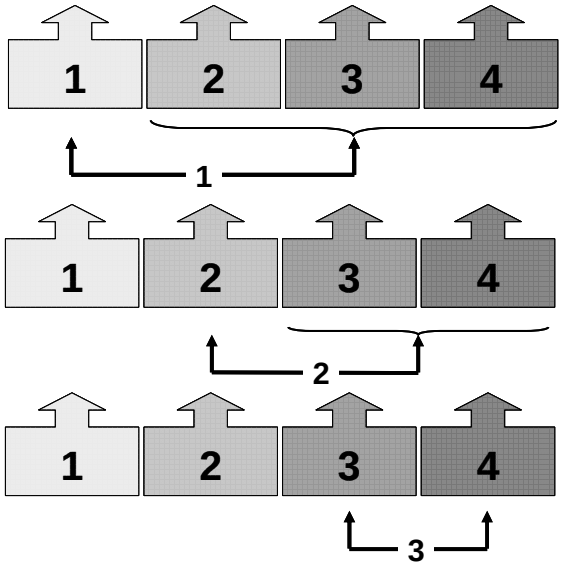
If there are q>1 of the p independent variables for which the proportional odds assumption fails, then for each of these q variables, a ß-parameter is estimated for each threshold. If q=p, then the model is equivalent to a generalised ordered logit model.

As with the POM, a major criticism of the PPOM has been the reliance on the need to see the dependent variable as being the result of grouping a latent variable in order to interpret the odds ratios. However, as discussed in relation to the POM, the model can be interpreted as a non-linear probability model, making it a flexible approach to modelling ordinal-level outcomes.

7.2.2.4 The continuation ratio model

The continuation ratio model (CRM) compares the probability of being in a particular category of the ordinal-level response, *j*, to being in a higher category (Figure 7.4), given that you have already reached category *j*.

Figure 7.4 Illustration of the comparisons made in a continuation ratio model: an example with four response categories



As with the POM, the effects of the independent variables are assumed to be the same at each level of the dependent variable. Unlike the POM however, this model is not invariant to the reversal or the collapsing of categories (Manor et al 2000, Lunt 2001b). This is because the CRM compares the probability of being in outcome category *j* to the probability of being in a category above *j* given that *j* has already been attained, thus discounting the data from categories below *j* (Box 7.4). The POM on the other hand uses all of the data in each comparison. This suggests that the CRM should be used when the categories of the dependent variable are themselves of interest (Ananth & Kleinbaum 1997, Manor et al 2000); the odds ratios produced represent "*the proportional change in the odds of being in any outcome category greater than* j *given that the subject is in the category* j *or above, when* x *changes by 1*" (Lunt 2005). This means that the CRM is generally thought to be most suitable when it is only possible to move between the categories of the dependent variable in one direction (Scott et al 1997). For example, when the outcome is

severity of radiographic osteoarthritis (none, mild, moderate, severe), and a return to a less advanced state is not thought possible.

Box 7.4 Mathematical description of the continuation ratio model

Let Y be an ordinal categorical dependent variable with k categories. Then there are k-1 categories from which it is possible to move to a higher category.

A binary logistic regression model uses a logit link function to model the probability that an individual is in each category. Then creating k-1 comparisons and fitting a binary logistic regression model to each gives,

$$\log_e \frac{P(y=1|x_i)}{P(Y \geqslant 2|x_i)} = \alpha_j + \beta_1 x_i , \quad \ldots\ldots, \quad \log_e \frac{P(Y=k-1)}{P(Y \geqslant k)} = \alpha_{k-1} + \beta_{k-1} x_i .$$

Interest is in the odds of progressing to a higher category, given that each category has been reached.

The CRM fits these equations simultaneously and assumes, as in the POM (Box 7.2), that $ß_1 = ß_2 = \ldots = ß_{k-1}$.

Hence, for a single independent variable, x, the CRM is modelled by the equation

$$\log_e \frac{P(Y=j|x_i)}{P(Y \geqslant j+1|x_i)} = \alpha_j + \beta x_i , \; j=1,\ldots, k-1,$$

where the $ß$-parameter does not depend on j, the level of the dependent variable.

In the case of p independent variables, this becomes

$$\log_e \frac{P(Y=j|x_i)}{P(Y \geqslant j+1|x_i)} = \alpha_j + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} , \; j=1,\ldots, k-1,$$

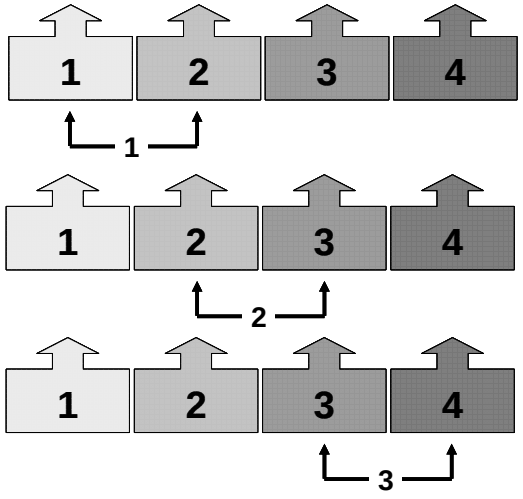
where again, the $ß$-parameter is independent of the level j, of Y.

7.2.2.5 The adjacent categories model

The adjacent categories model (ACM) compares the logits in each possible pair of adjacent categories simultaneously (O'Connell 2006: pg 76) (Figure 7.5). This model is useful when interest lies in what factors are associated with moving into the next highest category for adjacent pairs (O'Connell 2006: pg 76).

The ACM is subtly different from the CRM, where interest is in moving to any higher category than the present one (Box 7.5). The ACM uses less of the data in each comparison than the (P)POM or the CRM. As with the POM and the CRM, the effects of the independent variables are assumed to be the same at each level of the dependent variable.

Figure 7.5 Illustration of the comparisons made in an adjacent categories model: an example with four response categories



Box 7.5 Mathematical description of the adjacent categories model

Let Y be an ordinal categorical dependent variable with k categories. There are then k-1 possible adjacent comparisons, i.e. 1 vs 2, 2 vs 3,….,k-1 vs k. Let x_i be the independent variable x at level i.

A binary logistic regression model uses a logit link function to model the probability that an individual is in each category. Then creating k-1 pairs of adjacent categories and fitting a binary logistic regression model to each gives,

$$\log_e \frac{P(Y=1\mid x_i)}{P(Y=2\mid x_i)} = \alpha_1 + \beta_1 x_i, \quad \ldots\ldots, \quad \log_e \frac{P(Y=k-1\mid x_i)}{P(Y=k\mid x_i)} = \alpha_{k-1} + \beta_{k-1} x_i .$$

The ACM fits these equations simultaneously and assumes, as in the POM (Box 7.2) that $ß_1 = ß_2 = \ldots = ß_k$.

Hence, for a single independent variable, x_i, the ACM is modelled by the equation

$$\log \frac{P(Y=j\mid x_i)}{P(Y=j+1\mid x_i)} = \alpha_j + \beta x_i, \quad j=1,\ldots, k-1,$$

where the $ß$-parameter does not depend on j, the level of the dependent variable.

In the case of p independent variables, this becomes

$$\log \frac{P(Y=j\mid x_i)}{P(Y=j+1\mid x_i)} = \alpha_j + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \quad j=1,\ldots, k-1,$$

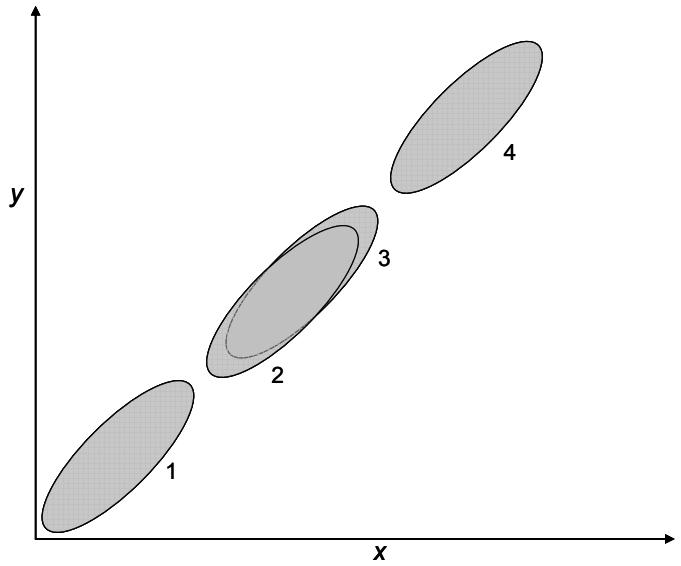
where again, the $ß$-parameter is independent of the level j, of Y.

7.2.2.6 The stereotype logistic regression model

The stereotype logistic regression model (SLRM) was developed by Anderson (1984) and is nested within the MLRM (Section 7.2.1.4), but has fewer parameters (Lunt 2001b). Lunt (2001b) describes the SLRM as a "*non-linear form of constrained multinomial model*".

Two key concepts are needed for the SLRM: distinguishability and dimensionality. Distinguishability is the ability of an independent variable to differentiate between two or more levels of the dependent variable (Figure 7.6). For example, if the categories of the dependent variable are "all days", "most days", "some days", "few days", "no days", there may be some confusion regarding which category from "some days" and "few days", represents more time, and this could lead to the categories essentially been viewed as one category. This property means that the SLRM is generally thought to be suitable for modelling dependent variables consisting of qualitative categories, where the ordering may not be clear with respect to the regression relationship (Anderson 1984). The case of using the SLRM for assessed responses is strengthen because the ordinality of the SLRM is not an assumption of the model, rather it is a part of the model building process (Lall et al 2002).

Figure 7.6 Distinguishability in a regression relationship. An example with a four response categories



Categories 1 and 4 of *y* are clearly distinguishable from the each other and from categories 2 and 3 with respect to *x*. Categories 2 and 3 of *y* are not distinguishable from each other with respect to *x*.

The dimensionality of a SLRM describes the number of functions or equations needed to distinguish between different levels of the dependent variable. In a unidimensional model, a single function of independent variables is required, whereas in a multidimensional model, multiple functions are needed. This can be illustrated using the example of severity of radiographic osteoarthritis (none, mild, moderate, severe) introduced earlier. If the same set of independent variables was associated with the onset of the disease and its progression, a single equation would describe the association. This is a unidimensional SLRM (Box 7.6). However, if different independent variables were associated with onset (movement from none to mild disease) and others were associated with progression (movement from mild through to moderate and severe disease), a second equation would be needed, and the model would be two-dimensional. The dimensionality of the model must be determined before it can be fitted, and if there is no prior information about this, this can be problematic. Interpretation may also be an issue in multidimensional models and some would argue, for example Lunt (2001b), that the comparisons made by multidimensional SLRMs mean that the model is not truly ordinal.

The SLRM is based on the MLRM and hence the comparisons made by the SLRM are not based on cumulative probabilities as in the other ordinal models discussed above. Instead they are like those of the MLRM where each category is compared to a baseline category (Figure 7.1). This has the potential to result in larger odds ratios than would be produced by the (P)POM, CRM or ACM, especially when comparisons involve extreme categories.

The SLRM is probably more flexible than any of the other ordinal models discussed in this section, but, it could also be considered to be the most complicated and most difficult to fit.

Box 7.6 Mathematical description of the unidimensional stereotype model

Let Y be an ordinal categorical dependent variable with k categories. Let x_i the independent variable x at level i.

For p independent variables, the probability of outcome j, given the values of the independent variable x, in the MLRM (Box 7.1) is given by,

$$\log_e \frac{P(Y = j \mid x)}{1 - P(Y = 1 \mid x)} = \alpha_j + \beta_j x \text{ , } j=1,\ldots,k.$$

Replacing β_js in the MLRM with $\varphi_j \beta_j$, the model equations for the SLRM are,

$$\log_e \frac{P(Y = j \mid x)}{1 - P(Y = 1 \mid x)} = \alpha_j + \varphi_j \beta_j x \text{ , } j=1,\ldots,k.$$

In order to identify the model, the following constraints are imposed,

$$\varphi_0 = 0,\ \varphi_k = 1,\ \beta_0 = 0.$$

For this to be an ordinal regression model, it is also necessary to impose the constraint, $\varphi_0 \leqslant \varphi_2 \leqslant \ldots \leqslant \varphi_k$. Alternatively, this constraint can be omitted and the ordinality of the model tested through the fitting process.

In the case of p independent variables, the model equation becomes,

$$\log_e \frac{P(Y = j \mid x)}{1 - P(Y = 1 \mid x)} = \alpha_j + \varphi_j \sum_{i=1}^{p} \beta_i x_i \qquad j=1,\ldots,k$$

7.3 Summary and discussion

As has been discussed in this chapter, there are several possible approaches to modelling ordinal-level dependent variables. The format of the data, and the specific research question will determine which model is most suitable. Before choosing an ordinal model for a particular purpose, there are two things that should be considered.

First, is the ordinality of the dependent variable important with respect to the independent variables? Anderson (1984) discusses that an apparently ordinal-level outcome does not necessarily imply an ordinal regression relationship; for this the dependent variable must be ordinal with respect to the independent variables.

Second, one must choose a suitable model. Given that one accepts the suitability of any type of ordinal regression model for qualitative categories or grouped continuous data, based on the concept of a non-linear probability model (Long & Freese 2006: pg 187), all of the models discussed in Section 7.2.2 are potential candidate models. If one does not accept this argument, response variables consisting of qualitative catgeories, such as individual PF-10 items, will require the SLRM or the MLRM. Assuming that one does not distinguish between models purely on the basis of the data generating process, all of the models in Section 7.2.2 are plausible.

The SLRM may be the most appropriate model if the ordinality of the dependent variable with respect to the independent variables is questionable. However, as this model makes all comparisons to a reference group, it makes less use of the data than some of the other models, and it has been argued that it is not truly ordinal (Lunt 2001b). For this reason, it should perhaps not be advocated except where ordinality is unclear or where it would be the simplest model; in the case of a large number of outcome categories.

Given that the ordinal-level nature of the dependent variable is established with respect to the independent variables, the POM provides an ordinal model that can be fitted by most statistical software packages. This model is possibly the easiest to interpret, and has featured in prominent epidemiological and medical journals (for example the American Journal of Epidemiology: Bhuiyan et al (2003), and Pain: Berglund et al (2006)). However, as discussed above, for the conclusions drawn from this model to be valid, the assumption of proportional odds must hold. Whether or not the assumption has been tested is rarely reported in journal articles and is frequently violated in practice.

The POM, and its extension for non-proportional odds, the PPOM, are probably best suited to situations where interest lies in associations between independent variables and the ordinal dependent variable as a whole, in particular when movement on the ordinal-level scale is thought to be possible in either direction. Where the dependent variable represents a progression, as in the radiographic osteoarthritis example used in this chapter, the POM may still be suitable, but a CRM is likely to be more appropriate as it models the probability of moving to a higher category, given you have already reached the current category. An ACM might be useful in situations where interest lies in factors associated with an increase of one category in the dependent variable. Again using the radiographic osteoarthritis example, interest might not be in progression to a higher category, but just to the next, adjacent category. The ACM allows this to be considered. In the case

of dependent variables with only three outcome levels, there is likely to be very little difference between the CRM and ACM and, to a certain extent, the POM. Differences between models will increase with increasing numbers of outcome categories.

Where there are a large number of outcome categories and the assumption of parallel logits fails for a POM, CRM or ACM, extending these models to allow for non-parallel lines may not be practical. In the case of a dependent variable with six outcome categories, the PPOM would produce five odds ratios for each independent variable that failed the assumption of proportional odds. In a case such as this, the SLRM may be a sensible option, as, unlike the PPOM, it does not require an odds ratio to be estimated for every level of the outcome for each independent variable if categories are indistinguishable with respect to the independent variable. However, the model with all categories distinguishable may be the best fitting.

Table 7.1 summarises some advantages and disadvantages of the models discussed in the previous sections and Table 7.2 details the comparisons of odds made by each model. Having considered the nature of the data and the research question, it is possible that more than one type of model will still be applicable to the situation. In these cases, the availability of computing power may need to be considered. How easy is access to software to fit each model, given that some packages allow greater flexibility than others? Should it be possible to fit two different types of models, Akaike's Information Criterion or a similar measure of model fit could be used to assess the relative suitability of each model. A major consideration here should probably be the interpretability of each of the models, and the number of parameters each produces. As is usually the case with statistical models, the more parsimonious model would probably be the best choice.

Despite the wide range of regression models available for ordinal dependent variables, some authors still question their usefulness. For example, Bender & Grouven (1998) have suggested that ordinal models are not necessary and that the use of a sliding dichotomy is easier in terms the checking of modelling assumptions. With a large number of outcome categories however, a sliding dichotomy is time consuming to fit and produces a large, and possibly unnecessary, number of model parameters. In the case of the PF-10 items considered in the next chapter, there are only three levels to the item response, but with a five- or even seven-level ordinal dependent variable, this approach would result in many odds ratios and would not be practical to implement or easy to interpret.

Table 7.1 Advantages and disadvantages of each approach to modelling ordinal data

Modelling approach	Advantages	Disadvantages
Ordinary least squares regression (Section 7.2.1.1)	Simple Easy to fit	Assumes highly unlikely variance structure
Binary logistic model (Section 7.2.1.2)	Simple Easy to fit	Ignores potential relationships between odds ratios at different thresholds
Sliding dichotomy (Section 7.2.1.3)	Uses all the data Easy to fit	Can produce a lot of parameter estimates Results can be repetitive Inefficient in terms of time and computing power
Multinomial logistic model (Section 7.2.1.4)	Makes few assumptions	Lots of parameter estimates Does not account for ordering
Generalised ordered logit model (Section 7.2.2.1)	Uses all the data More efficient than sliding dichotomy	Can produce a lot of parameter estimates Results can be repetitive
Proportional odds model (Section 7.2.2.2)	Simplest truly ordinal model Easy to fit	Strong assumption Arguably unsuitable for assessed outcomes
Partial proportional odds model (Section 7.2.2.3)	More flexible than POM More parsimonious than the MLRM	Not able to be fitted in all statistical packages Can estimate negative probabilities Arguably unsuitable for assessed outcomes
Continuation ratio model (Section 7.2.2.4)	Ideal for when dependent variable represents a progression	Strong assumptions Does not use all of the data Not able to be fitted in all statistical packages
Adjacent categories model (Section 7.2.2.5)	Ideal for modelling increases of one category in the dependent variable	Strong assumptions Does not use all of the data Not able to be fitted in all statistical packages
Stereotype logistic model (Section 7.2.2.6)	Does not assume ordering but can allow for it if necessary Always considered suitable for an assessed response	Not able to be fitted in all statistical packages Difficult to choose dimensionality Does not use all of data Questionable ordinality

Table 7.2 Comparisons made in each ordinal modelling approach: an example with four response categories

Models	Comparisons	
	Reference	Comparison
Multinomial logistic model	1	2
Stereotype model	1	3
	1	4
Sliding dichotomy	1	2,3 4
Generalised ordered logit model	1,2	3, 4
(Partial) Proportional odds model	1, 2 ,3	4
Continuation ratio model	1	2, 3, 4
	2	3, 4
	3	4
Adjacent categories model	1	2
	2	3
	3	4

In general, arguments that ordinal model are more difficult to interpret that the binary logistic regression model are easily refuted, because all of the ordinal models presented in this chapter are based on the binary logistic regression model and so interpretation is rarely markedly different than for this binary model. Model fitting and interpretation should not remain a reason for favouring a binary model over an ordinal model where fuller use could be made of the data.

Furthermore, a sliding dichotomy could only be considered an alternative to a POM or PPOM, and as has been discussed in this chapter, other models make different comparisons that may be more appropriate in different situations. The choice of models that allow different comparisons between the categories of the ordinal dependent variable to be made allows far greater flexibility in analysis than a binary model. That is not to say that data should be analysed and reanalysed until the preferred result is found, but different ordinal models can be used in different situations.

In practice there may not be major differences between binary and ordinal models for particular datasets (Manor et al 2000). However, it is not possible to know this without choosing and fitting an appropriate ordinal model with which to compare the binary model. Ordinal models should be used as a point of principle. They are available and make best use of the data. There is

therefore no reason that these models should be ignored when they offer the most appropriate means of analysis. For this reason, the next chapter considers the ordinal modelling options available and assesses the association between socio-demographic variables and pain, and locomotor disability, as assessed by the individual PF-10 variables.

8 Modelling the cross-sectional association of the PF-10 locomotor disability-specific items with socio-demographic factors and pain in the NorStOP

8.1 Introduction

Chapter 6 identified five items from the PF-10 as being likely indicators of locomotor disability. Analyses in the NorStOP datasets and a review of the literature showed them to have acceptable repeatability over time and provided evidence of their validity for this purpose. Although these items do not constitute measures of locomotor disability as defined in Chapter 2, they potentially provide a more fine-grained approach to the assessment of locomotor disability in large-scale epidemiological surveys than has usually been the case (see Chapter 3).

The majority of the previous literature regarding the epidemiology of locomotor disability has used binary definitions of disability, considering study participants to be either 'disabled' or 'non-disabled'. In particular, many studies have examined the association between socio-demographic factors and pain on the presence or onset of locomotor disability. This chapter will consider the approaches to regression modelling described in Chapter 7 and using the most appropriate methodology will assess the association between these factors and locomotor disability cross-sectionally in the NorStOP cohort at baseline. The aim of these analyses is to investigate the potential advantages of an ordinal-level outcome over a binary outcome, in such association studies.

8.2 Methods

8.2.1 Dependent variable

The analyses in this section used one of the locomotor disability-specific PF-10 items: Item i (walk 100 yards). This item was chosen as it logically represents one of the most severe forms of locomotor disability within the instrument; limitation in climbing one flight of stairs being the other. Furthermore, this PF-10 item represents a task necessary for many everyday activities, for example walking to a car or around a local shop. In addition, this item has been less frequently

criticised than other PF-10 items (for example, Mallinson (2002)). Analyses of the other PF-10 locomotor disability-specific items at baseline are presented in Appendix D. The results of these analyses and their relationship to the item concerning walking 100 yards are considered in Section 8.3.

8.2.2 Independent variables

This section describes how the independent variables of interest, i.e. socio-demographic factors and pain, were defined and used in the analyses presented in this chapter. It was chosen to model these associations in particular, as they have been consistently shown to be associated with the onset of locomotor disability in previous studies (Ahacic et al 2000 and 2003, Clark et al 1998 and 1998b, Ebrahim et al 2000, Gill et al 2006, Guralnik et al 1993 and 2001, Leveille et al 2000 and 2007, Nordstrom et al 2007, Sainio et al 2007, Wannamethee et al 2005).

8.2.2.1 Demographic characteristics and socioeconomic status

Gender, age and socioeconomic status were self-reported at baseline. Age-groupings in the regression models were in five-year bands from 50 to 54 years up to 90 years and over. Age-groupings for stratification were in 10-year bands from 50 to 59 years to 80 years and over. As described in Chapter 4, three indicators of socioeconomic status were used: educational attainment (school-age only versus further education), occupational class (Office for National Statistics 2000 and 2002) (manual: lower supervisory/technical, semi-routine occupations, routine occupations; versus non-manual: higher managerial, higher professional, lower managerial/professional, intermediate occupations; self-employed people were excluded) and perceived adequacy of income (Thomas 1999) (inadequate: “find it a strain to get by from week to week”, “*have to be careful with money*”; versus adequate: “Able to manage without much difficulty”, “*quite comfortably off*”). Living arrangement was self-reported as living alone or not.

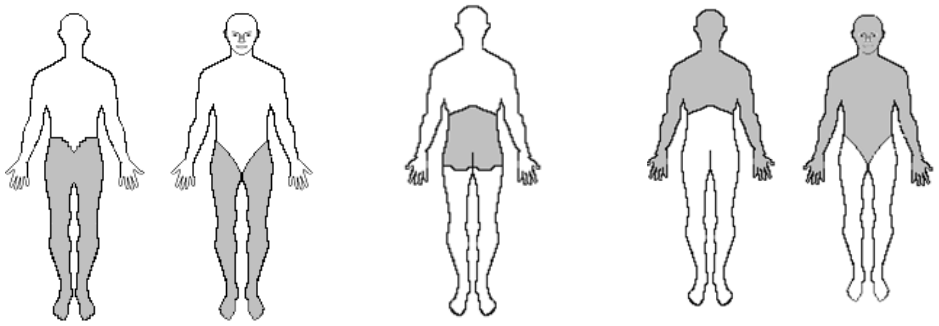
8.2.2.2 Pain

In this chapter, pain was defined as described in Chapter 4. Briefly, those individuals who responded “*Yes*” to the pain screening item and shaded the manikin were considered to have pain

in the shaded area(s). Those who responded "*No*" to the screening item and did not shade the manikin were considered to have no pain. All other respondents were excluded from the pain analyses.

The manikin was used to define 'lower limb pain' (including the hips (Birrell et al 2001)) and 'low back pain' (Papageorgiou et al 1995). Pain outside these regions was defined as 'pain elsewhere' (Figure 8.1). These three pain locations were then used to define a six-level categorical variable for pain location: 'no pain', 'pain not the lower limb', 'pain in the lower limb only', 'pain in the lower limb and low back', 'pain in the lower limb and elsewhere', and 'multiple' pains, i.e. pain in all three areas.

Figure 8.1 Definitions of lower limb pain, low back pain and pain elsewhere



8.2.3 Aims of analyses

These analyses had two main aims. First, to assess the association between pain and locomotor disability independently of gender, age, educational attainment, occupational class, perceived adequacy of income and living arrangement. Second, to assess the role of gender, age, educational attainment, occupational class, and perceived adequacy of income as effect modifiers in the association between pain and locomotor disability.

8.2.4 Statistical analyses

8.2.4.1 Appropriate choice of ordinal model

Several models for ordinal dependent variables were presented in Chapter 7. It was necessary to choose the most appropriate type of model to answer the particular research question in these data. This section describes how this choice of model was made.

PF-10 items are assessed response variables and although it is reasonable to assume that there is an underlying continuum of locomotor disability, the categories of the item responses are unlikely to have resulted directly from this continuum. Respondents will have had to draw on several pieces of information in order to assign themselves to a level of limitation, for example 'how far is 100 yards', 'can I walk this far?', 'if I can, how limited would I be?'.

Accepting the conceptualisation of non-linear probability models suggested by Long and Freese (2006; pg 187), all of the models discussed in Chapter 7 were available for use in these analyses. However, it is statistically desirable to use a model derived specifically for ordinal data, rather than one of the non-ordinal options discussed in Chapter 7.

The CRM was considered unsuitable for use in modelling locomotor disability, as it implies that return to a lower level of disability is not possible. The literature review in Chapter 3 showed that this is not the case when considering binary definitions of locomotor disability (Wannamethee et al 2005, Gill et al 2006), and there is no reason to suggest that this would not be the case when considering an ordinal assessment of disability.

The ACM was also considered unsuitable for these analyses, as it models only the probability of being in the next category of the dependent variable. It has been shown that, although locomotor disability usually develops progressively, as would be modelled by the ACM, onset can also be catastrophic, where participants move from no disability to severe disability without going through the intermediate stages (Guralnik et al 2001).

This leaves the POM, PPOM and the SLRM. As discussed in Chapter 7, the SLRM compares each category to a reference group. This again suggests that each category is of interest and makes the ordinality of the model questionable. Furthermore, in this case, the dependent variable has only three categories, meaning that the potential parsimony of the SLRM over the PPOM is unnecessary. The two remaining ordinal models were therefore the POM and the PPOM.

The POM would be favoured for its simplicity in comparison to the PPOM, but its assumption of proportional odds is very strong. Therefore, a PPOM was considered to be the most appropriate model in this case. Should the assumption of proportionality for any independent variable be met, a single odds ratio would represent the effect of this independent variable at all levels of the dependent variable. In the event that this occurred for all independent variables, the PPOM would simplify to the POM.

8.2.4.2 Model fitting

In order to meet the aims of this chapter, responses to the PF-10 Item i (walk 100 yards) were tabulated against each of gender, age, educational attainment, occupational class, perceived adequacy of income, living arrangement and pain category to assess the extent of any ordinal association. Assuming an ordinal response, a PPOM was then fitted in Stata 9.2, using the user written program -gologit2- (Williams 2006). The dependent variable was PF-10 Item i (walk 100 yards). Independent variables were pain category, gender, age, educational attainment, occupational class, perceived adequacy of income and living arrangement. Likelihood ratio tests were used in a stepwise selection procedure to test for non-proportionality of the independent variables. All independent variables in all PPOMs that were modelled as proportional in the final model, i.e. only one odds ratio was calculated, therefore conformed to the assumption of proportional odds.

To assess the role of gender, age, educational attainment, occupational class and perceived adequacy of income as effect modifiers, the PPOM was repeated stratified by each of these factors. Age was split into 10-year bands for this stratified analysis.

As with most regression models, the PPOM requires complete data on any individual, in order for that individual to be included in a model. Therefore, persons were excluded from the analysis if they did not have data on the dependent variable, i.e. PF-10 Item i (walk 100 yards), or one of more of the independent variables. Results are presented accompanied by the actual number of people included in the final model. All results are presented as odd ratio (OR), with accompanying 95% confidence interval (CI) unless otherwise stated.

All of these analyses for PF-10 Item i (walk 100 yards) were then repeated for the four remaining locomotor disability-specific PF-10 items.

8.3 Results

Results of the analyses of the PF-10 item relating to walking 100 yards are described in this section. Analyses of the other four PF-10 locomotor-related items are shown in Appendix D.

8.3.1 Crude associations of pain and socio-demographic characteristics with locomotor disability

Cross-tabulations of pain and the socio-demographic characteristics outlined above with the level of limitation in walking 100 yards are given in Table 8.1. Female gender, older age, lower socioeconomic status (no further education, manual occupation, perceived inadequate income), living alone and the presence of pain (particularly pain at multiple sites) was associated with a greater proportion of people with either a little or a lot of limitation in walking 100 yards. There were clear ordinal associations between pain and the socio-demographic characteristics, and locomotor disability, suggesting that an ordinal modelling approach was suitable for these data.

Table 8.1 Distribution of limitation in walking 100 yards by the socio-demographic characteristics and pain category. NorStOP, n=17,957

n (%)	Level of limitation in walking 100 yards		
	A lot	A little	None
Overall	2,206 (12)	3,045 (17)	12,706 (71)
Gender			
Males	856 (11)	1,164 (15)	5,960 (75)
Females	1,350 (14)	1,881 (19)	6,746 (68)
Age-group (years)			
50 to 54	132 (5)	227 (9)	2,174 (86)
55 to 59	231 (7)	366 (11)	2,732 (82)
60 to 64	252 (9)	391 (14)	2,161 (77)
65 to 69	286 (10)	456 (17)	2,026 (73)
70 to 74	318 (13)	489 (20)	1,629 (67)
75 to 79	318 (16)	513 (26)	1,161 (58)
80 to 84	357 (27)	372 (28)	587 (45)
85 to 89	200 (36)	174 (31)	186 (33)
90 and over	112 (51)	57 (26)	50 (23)
Educational attainment[a]			
Further Education	145 (7)	235 (11)	1,706 (82)
School	1,980 (13)	2,729 (18)	10,756 (70)
Occupational class[a]			
Non-manual	375 (8)	600 (12)	3,952 (80)
Manual	1,335 (13)	1,905 (18)	7,177 (69)
Adequacy of income[a]			
Adequate	831 (9)	1,292 (13)	7,638 (78)
Inadequate	1,270 (16)	1,660 (21)	4,847 (62)
Living arrangement[a]			
Does not live alone	1,322 (10)	1,869 (15)	9,650 (75)
Lives alone	741 (18)	971 (23)	2,497 (59)
Pain category[a]			
No pain	184 (4)	347 (8)	4,103 (89)
Pain not in the lower limb	110 (6)	184 (10)	1,512 (84)
Lower limb pain only	199 (13)	285 (19)	994 (67)
Pain in lower limb and lower back only	178 (12)	230 (16)	10,58 (72)
Pain in lower limb and elsewhere only	351 (14)	527 (22)	1,560 (64)
Multiple-pains	864 (21)	1,073 (26)	2,248 (54)

[a]Subject to missing data

8.3.2 Adjusted associations of socio-demographic characteristics and pain with locomotor disability

The model fitting process for the PPOM indicated that the assumption of proportional odds was violated for the independent variables gender and living arrangement, i.e. these variables have a different effect at each level of locomotor disability. All other independent variables met the assumption of proportional odds. Table 8.2 shows the adjusted odds ratios, with 95% confidence intervals, from the fitting of this model. All independent variables were significantly associated with limitation in walking 100 yards at one or both thresholds of disability.

Females were more likely to have any limitation in walking 100 yards than were males (adjusted OR 1.13; 95% CI 1.03, 1.24). However, gender did not distinguish those with a lot of limitation from those with no or a little limitation (0.97; 0.86, 1.10).

Older age was associated with an exponential rise in limitation at both thresholds, i.e. any limitation compared to no limitation, and a lot of limitation compared to no or a little limitation. The strength of this association increased more rapidly after the age of 75 years. Those who were aged 85 to 89 years had nearly 15 times the odds of being more limited than a similar person aged 50 to 54 years (14.66; 11.26, 19.10).

Those in lower socioeconomic groups had higher levels of locomotor disability, however socioeconomic status was defined. The effect of all socioeconomic variables was the same at each level of limitation. Similarly to gender, living alone was associated with any limitation (1.20; 1.08, 1.34), but not with the level of limitation (0.97; 0.85, 1.12).

Table 8.2 The adjusted association of pain and socio-demographic characteristics with locomotor disability (limitation in walking 100 yards). PPOM. NorStOP, n=12,882

Odds ratio (95% confidence interval)	"Yes, limited a little" or "Yes, limited a lot" [b]	"Yes, limited a lot" [c]
Gender		
Male	1	1
Female	1.13 (1.03, 1.24)	0.97 (0.86, 1.10)
Age-group (years)		
50 to 54 years	1	
55 to 59 years	1.23 (1.03, 1.47)	
60 to 64 years	1.76 (1.48, 2.09)	
65 to 69 years	2.17 (1.82, 2.58)	
70 to 74 years	3.28 (2.75, 3.91)	
75 to 79 years	4.80 (4.00, 5.75)	
80 to 84 years	8.73 (7.15, 10.65)	
85 to 89 years	14.66 (11.26, 19.10)	
90 years and over	38.25 (25.34, 57.72)	
Educational status[a]		
Further education	1	
School age education	1.27 (1.09, 1.49)	
Occupational class[a]		
Non-manual occupation	1	
Manual occupation	1.42 (1.28, 1.56)	
Adequacy of income[a]		
Adequate	1	
Inadequate	2.00 (1.83, 2.19)	
Living arrangement[a]		
Does not live alone	1	1
Lives alone	1.20 (1.08, 1.34)	0.97 (0.85, 1.12)
Pain category[a]		
No pain	1	
Pain not in the lower limb	1.82 (1.51, 2.19)	
Lower limb pain only	3.69 (3.10, 4.38)	
Lower limb pain and lower back pain only	3.40 (2.85, 4.07)	
Lower limb pain and pain elsewhere only	4.59 (3.96, 5.33)	
Multiple-pains	8.05 (7.04, 9.20)	

[a]Subject to missing data; [b]Reference category is "No, not limited at all"; [c]Reference category is "No, not limited at all and "Yes, limited a little"

Pain was significantly associated with limitation, with all pain groups having increased odds of limitation compared to those without pain. Those with pain in the lower limb only had approximately four times the odds of being limited than a similar person with no pain (3.69; 3.10, 4.38). The addition of low back pain to lower limb pain had little effect on the association with limitation. However, lower limb pain and pain elsewhere only (4.59; 3.96, 5.33) and multiple-pains (8.05; 7.04, 9.20), again increased the odds of limitation.

Similar analyses of the other locomotor disability-specific PF-10 items showed comparable associations, although the proportionality of some independent variables differed between items (Appendix D). In the analyses of the other items, pain and the socio-demographic characteristics were significantly associated with locomotor disability at both dichotomies. The main point of interest from these analyses was a comparison of the magnitude of odds ratios between the items. Within the items relating to walking and to climbing stairs separately, there was an increasingly strong association in odds between gender and limitation, as the tasks became more difficult. For example, the odds ratio associated with gender for climbing several flights of stairs was 1.8, whilst for climbing one flight it was 1.5. Within the walking questions, the effect of older age was strongest in the more difficult tasks, with the gap between tasks widening with older age. The effect of socioeconomic status was similar across task difficulties. In the walking items, pain had a similar magnitude of effect across difficulty levels, whilst in the stair-climbing items, there was a stronger association between pain and limitation in climbing a single flight of stairs than between pain and limitation in climbing several flights of stairs.

8.3.3 Stratified, adjusted associations between pain and locomotor disability

Stratified multivariable analyses showed that there were no differences in the effect of pain between those with school and further education, manual and non-manual occupations and perceived inadequate and adequate incomes (Appendix D).

An increase in odds ratios with increasing pain was seen for both genders, although this was more pronounced in females (Table 8.3). For females, multiple-pains conferred a nine-fold increase in the odds of limitation (9.20; 7.67, 11.03) compared to no pain, whilst in males; multiple-pains conferred a seven-fold increase (6.74; 5.52, 8.23).

Table 8.3 The adjusted association of pain[a] and locomotor disability (limitation in walking 100 yards). PPOM stratified by gender. NorStOP, n=12,882

Odds ratio (95% confidence interval)	Proportional odds ratio[b]	
	Male, n=5,723	Female, n=7,159
No pain	1	1
Pain not in the lower limb	1.80 (1.37, 2.36)	1.82 (1.41, 2.35)
Lower limb pain	2.96 (2.29, 3.83)	4.45 (3.52, 5.62)
Lower limb pain and lower back pain only	2.55 (1.94, 3.34)	4.23 (3.34, 5.36)
Lower limb pain and pain elsewhere only	3.93 (3.14, 4.92)	5.19 (4.25, 6.34)
Multiple-pains	6.74 (5.52, 8.23	9.20 (7.67, 11.03)

[a]Subject to missing data; [b]Model is adjusted for age, educational attainment, occupational class, perceived adequacy of income and living arrangement

There was the smaller effect of multiple-pains at older ages (Table 8.4) with multiple-pains conferring an 11-fold increase in the odds of limitation (10.69; 7.75, 14.73) compared to no pain, in the 50 to 59 year age-group but only a five-fold increase in those aged 80 years and over (5.10; 3.72, 7.00). The effects of other pain groups were similar across the age-groups.

In similar analyses of the remaining four questions, similar patterns were found (Appendix D). The stronger effect of pain in females than in males and the decreased effect of pain at older ages were replicated.

Table 8.4 The adjusted association of pain[a] and locomotor disability (limitation in walking 100 yards). PPOM stratified by age-group. NorStOP, n=12,882

Odds ratio (95% confidence interval)	Proportional odds ratio[b]			
	50 to 50 years, n=4,592	60 to 69 years, n=4,119	70 to 79 years, n=2,972	80 years and over, n=1,199
No pain	1	1	1	1
Pain not in the lower limb	1.88 (1.21, 2.94)	1.52 (1.04, 2.20)	2.10 (1.53, 2.88)	1.92 (1.25, 2.95)
Lower limb pain	3.62 (2.34, 5.60)	3.73 (2.60, 5.23)	4.02 (3.01, 5.39)	3.36 (2.32, 4.87)
Lower limb pain and lower back pain only	4.18 (2.75, 6.34)	3.18 (2.27, 4.46)	3.06 (2.24, 4.18)	3.23 (2.15, 4.84)
Lower limb pain and pain elsewhere only	4.78 (3.12, 6.89)	4.94 (3.70, 6.59)	4.59 (3.56, 5.91)	3.90 (2.80, 5.44)
Multiple-pains	10.69 (7.75, 14.73)	8.67 (6.68, 11.25)	6.63 (5.26, 8.35)	5.10 (3.72, 7.00)

[a]Subject to missing data; [b]Model is adjusted gender, educational attainment, occupational class, perceived adequacy of income and living arrangement

8.4 Discussion

8.4.1 Principal findings

The analyses in this chapter have shown that cross-sectionally, there is an ordinal association of pain, and the socio-demographic characteristics of gender, age, socioeconomic status and living arrangement with the level of locomotor disability. The association with pain was strongest when a person had multiple-pains, i.e. pain in the lower limb, low back and elsewhere, with the effect of these pains being greater than the sum effects of the individual pains. This is in agreement with, and extends, the results of Melzer et al (2005) and Croft et al (2005), who found that pain is common in middle- and old-age and that the presence of pain at multiple sites is associated with higher levels of disability than single-site pain. The proportionality of the associations varied across the locomotor disability-specific PF-10 items used as dependent variables in these analyses, although pain and all socio-demographic characteristics remained significantly associated with locomotor disability whichever item was used to assess this.

Stratified analyses showed that the effect of pain was similar in different socioeconomic groups. However, the association between multiple-pains and locomotor disability, adjusted for age, socioeconomic status and living arrangement was stronger in females than in males. This is in contrast to the findings of Odding et al (1995), who showed that the number of joints affected by osteoarthritis was more strongly related to locomotor disability in males than in females. This perhaps suggests that the pain reported in the current analyses is not necessarily joint pain and may include other causes of pain, for example cardiovascular disease or neurological conditions. These causes of pain may have a larger effect on disability in females than in males.

The association between multiple-pains and locomotor disability was less strong at older ages when examined within 10-year age bands. This is in agreement with the findings of Melzer et al (2005) concerning the association of pain on walking with locomotor disability. It also supports the findings of Odding et al (1995) who showed a similar pattern in adults aged 55 years and over in the Netherlands.

8.4.2 Strengths and weakness of the study

These analyses have used a large sample size to show strong associations between specific independent variables and locomotor disability, using the locomotor disability-specific PF-10 items. However, as all of these analyses were carried out using cross-sectional data, the scope of the conclusions that can be drawn here is limited: it is difficult to disentangle the order in which, say, pain and locomotor disability are developed and it is not possible to infer causal associations. For this, prospective data are needed, and this will be considered further in Chapter 14.

The large sample size did allow for the assessment of the association between pain and locomotor disability in different population groups. This would not have been possible with a smaller sample size, and thus, findings such as the stronger association between pain and locomotor disability in the younger age-groups than in the older age-groups, would have been missed.

A very stringent rule was employed in the definition of pain. As described in Chapter 4, those people defined to have no pain had responded "*No*" to the pain screening item and had not shaded the manikin, those people defined to have pain had responded "*Yes*" to the screening item and shaded the manikin. This resulted in a clear definition of pain, but also in 2,139 (11.6%) people being excluded from the analysis. This could have resulted in a bias if those people who did not complete the item and manikin in accordance with each other showed a different association between pain and locomotor disability than those who completed the item and manikin as required by this definition. This seems unlikely and is not considered to represent a major source of bias. Due to the large number of people lost from the analysis using this method to define pain groups, a different method is used in Chapter 14, as is described in Chapter 4.

The use of five different locomotor disability-specific items from the PF-10 is both a strength and weakness of this study. This approach allowed the association between pain and locomotor disability to be tested using a number of different 'definitions' of disability, i.e. limitations in different amounts of walking and stair-climbing. Using this single-item approach, it was also clear in which activities a person had difficulty, something that is not the case when disability is defined as a problem in one or more of a set of locomotor tasks (see Chapter 3). However, this approach produced rather unwieldy and repetitive results, with the same analyses essentially being

presented five times, once for each item. If a method could be developed to combine these items into a single measure of locomotor disability, as suggested in Chapter 6, this would be preferable. Chapters 9 to 12 address this issue in full.

8.4.3 Strengths and weaknesses of the study in relation to the current literature

Melzer et al (2005) have previously shown, in a similarly aged population, that pain on walking is associated with locomotor disability. The current analyses extend these findings to pain in general, rather than pain on activity. Although the results of these analyses confirm an association between locomotor disability and lower limb pain, the exponential rise in the odds of locomotor disability with age, regardless of pain status, suggests that lower limb pain is not solely responsible for locomotor disability. Further investigation is needed into potential determinants of locomotor disability in a longitudinal setting.

Ordinal models have been discussed in the statistical literature for over 20 years (for example McCullagh & Nelder (1983) and Anderson (1984)). However, they have been used infrequently in epidemiological research (Forrest & Andersen 1986, Ananth & Kleinbaum 1997). Often in this field, ordinal dependent variables are dichotomised and analysed using binary logistic regression or similar binary modelling techniques (for example Avlund et al 2003). As discussed in Chapter 7, this can lead to a loss of statistical information, power and efficiency and the placement of the dichotomy is often arbitrary, again losing information. The main reason for the frequent dichotomisation of ordinal dependent variables seems to be that binary variables are perceived to be easier to analyse, and the resulting parameter estimates easier to interpret. It is true that whilst all major statistical packages provide the opportunity to analyse binary dependent variables, fewer have the ability to analyse ordinal outcomes. However, this is changing. Most standard packages now provide functions with which to fit at least basic ordinal models, such as the POM. Some packages, for example Stata 11 (StataCorp LP 2009), go considerably further and allow all of the models described in this chapter to be implemented, although this may be through user written programs rather than the standard components of the packages.

Aside of the practical issues of analysis and interpretation, some authors, for example Bender & Grouven (1998) have suggested that ordinal models are not necessary and that the use

of a sliding dichotomy is easier in terms the checking of modelling assumptions. Whilst this is true, it should still be argued that with a large number of outcome categories, a sliding dichotomy may be time consuming to fit and produces a large, and possibly unnecessary, number of model parameters. In the case of the analyses presented in this chapter, there were only three levels to the item response, but because the five locomotor disability-specific items were considered separately, this approach would have resulted in many odds ratios to interpret and would not have been practical. Also, a sliding dichotomy could only be considered an alternative to a POM or PPOM, and as discussed in Chapter 7, other models make different comparisons that may be more appropriate in different situations.

This chapter has shown that it is not necessary to dichotomise ordinal dependent variables (whether at a single point or several), and much greater use can be made of these ordinal data using the appropriate statistical methodology. Here the use of the PPOM did not force a dichotomy to be chosen; hence it did not make assumptions about the important comparisons to make in terms of levels of disability, when there were no a priori hypotheses about this. Also, it allowed associations to differ between the possible dichotomies and permitted the investigation of potential differences. In the case of limitations in walking 100 yards, it was possible, using the PPOM, to infer that gender is only significantly associated with the presence of locomotor disability and not with its severity. Although the effect of non-proportionality in gender may not be substantial in this case, the results highlight the potential for ordinal modelling techniques to provide a wealth of information not available through the use of binary models. In addition, in this example it may give some insight into the different ways in which males and females choose to answer survey questions: do males prefer to put themselves at the extreme of the scale whilst females are more comfortable with the response options in the middle of the scale?

8.4.5 Meaning of the study

The fact that pain is more threatening to locomotion in middle-age than in the oldest age-groups, would suggest that the younger age-groups should be targeted for pain reduction therapies as well as the oldest groups. In the oldest age-groups, the proportion of people with locomotor disability is high, even in those without pain. However, the reduced odds ratios in these groups do not appear to be simply an artefact of the mathematics of the model. One explanation for the stronger association in the younger age-groups might be the causes of the pain. It could be that pain, in particular pain at multiple sites, in the youngest age-groups is caused by conditions such as fibromyalgia, whilst in the oldest age-groups, it is caused by generalised osteoarthritis. These different causes of pain may alter the effect that pain has on locomotion.

Having accounted for the effect of pain in these analyses, older age was still significantly associated with locomotor disability. This suggests that there is something else, other than pain, causing the increase in locomotor disability with age. Indeed, the proportion of people with pain is fairly stable across age-groups. It would seem likely that other morbidities are also affecting the locomotion of older adults. Previous studies, for example Kriegsman et al (1997) and Melzer et al (2005) have shown that, amongst other things, cardiovascular disease, vision problems and mental health problems are associated with locomotor disability in older adults. In the current analyses, pain in areas other than the lower limb was also associated with a higher likelihood of locomotor disability compared to those with no pain and the effect of pain elsewhere in the body, i.e. outside the lower limb and low back, added to lower limb pain has a greater effect than pain in the low back added to pain in the lower limb. These findings suggest that non-musculoskeletal pain may be related to this locomotor disability. For example, pain in the chest could be a result of angina, although it is not necessarily the pain, rather the other symptoms of the angina, such as breathlessness, that are associated with the locomotor disability.

These analyses have shown that pain is significantly associated with locomotor disability, but that the association is less strong at the oldest ages. However, gender, age and socioeconomic status are also independently associated with locomotor disability. This suggests that interventions aimed at reducing pain should be targeted not only at the younger age-groups but also specifically at females and those in socially disadvantaged groups.

Although it is not possible to draw absolute conclusions from these cross-sectional analyses, the data strongly suggest that, in most people, the path to locomotor disability is progressive, rather than catastrophic. This is inferred from the decreasing proportion of people with no limitation and increasing proportion with a lot of limitation through the age-groups, whilst the proportion with a little limitation stays relatively constant, thus supporting the findings of Guralnik et al (2001). Furthermore, these analyses showed higher levels of limitation in walking longer distances and in climbing more flights of stairs than in walking shorter distances and climbing fewer stairs. However, the results of these analyses do not imply that progression is inevitable or improvement impossible. Wannamethee et al (2005) and Gill et al (2006) have shown that recovery from and improvement in locomotor disability is possible and the use of the PPOM in these analyses has allowed for this.

As presented in Chapter 7, several different ordinal models are available and these can allow a number of different comparisons to be made using the ordinal data; allowing far greater flexibility in analysis than a binary model. That is not to say that data should be analysed and reanalysed until the preferred result is found, but different ordinal models can be used in different situations. In these analyses it was considered appropriate to use an ordinal model that used the whole of the dataset in each comparison and allowed for theoretical recovery from locomotor disability. Again, the cross-sectional nature of the data means that there was no recovery, but in theory, it is believed that recovery is possible and other studies have shown this, for example Wannamethee et al (2005) and Gill et al (2006).

8.4.4 Unanswered questions and future research

Whilst the analyses presented in this chapter have undoubtedly provided evidence regarding the cross-sectional association between pain and locomotor disability and also shown that the level of locomotor disability is higher at older ages, there are many questions that remain to be answered. In particular, Chapter 14 will address the longitudinal association between pain and locomotor disability, including the role of the onset of and recovery from pain.

As the cross-sectional analyses presented here cannot disentangle the potential causal effect of pain on locomotor disability; it is also possible that disability has a causal effect on pain.

Although this seems unlikely, it is possible that disability occurs first and alters the mechanism of locomotion, resulting in pain.

Furthermore, the association between pain, age and locomotor disability should be further considered. These analyses have provided evidence that the association between pain and locomotor disability is stronger at younger ages, whist the overall level of disability, in terms of the proportion of people reporting any limitation, increases with age. This is driven by the proportion of people reporting “a lot” of limitation. What is not clear from the analyses conducted in this chapter is the reason for this exponential increase with age. This increase does not seem to be entirely pain related, despite the obvious association between pain and locomotor disability, even at the oldest ages. Further investigation in this area would doubtless prove useful in understanding these associations.

Having to consider each of the PF-10 locomotor disability-specific items individually is unwieldy and it would be useful to combine these items into a single measure of locomotor disability. Aside of the increased practicality of having a single, interval-level measure, this would also allow the future investigations suggested above to be carried out to the fullest extent, considering a quantification of disability, as well as its presence or absence. This was discussed in Chapter 6 and it was concluded that the commonly used methods of summated ratings, i.e. assign each response option a value and sum the values across items, is not an appropriate method for creating an interval-level measure. Further work is necessary to assess possible methods for the combination of items into a measure. Work in this area is presented in the following chapters.

8.5 Conclusions

Although ordinal regression modelling is clearly a useful method for the analysis of individual dependent variables on an ordinal-level, most studies, particularly in epidemiology, collect data on many aspects of an individual, and ordinal models can only analyse one aspect at a time. This leads to a loss of information and efficiency. In the data used in this chapter, five items were analysed separately. This led to lengthy tables and discussion around the findings. However, there is evidence from the differing levels of limitation reported on each item (see Chapter 6), and from the varying strengths of association between the locomotor disability-specific items and the

independent variables, that there may be an ordering to the difficulty of the items. Hence there is possibly an order in which people develop the different locomotor disabilities considered here. Furthermore, these ordinal-level item responses do not represent true measurement of locomotor disability, as described in Chapter 2 and do not allow the continuum of disability discussed by Thomas (2007) to be fully investigated.

The next chapter describes the Rasch unidimensional measurement model (Rasch 1960), which has been mathematically proven to be the only method by which ordinal items responses can be transformed into an interval-level measure (Fischer 1995). In Chapter 10 this technique is used to derive a scoring mechanism for the five locomotor disability-specific items of the PF-10 into a single, interval-level measure of locomotor disability in the NorStOP pilot dataset. Following this, Chapter 11 assesses the generalisability of the scoring mechanism to other datasets and Chapter 12 assesses the psychometric properties of the interval-level measure.

9 The Rasch unidimensional measurement model: concepts and theory

9.1 Introduction

In Chapter 8, the five PF-10 locomotor disability-specific items (d - climb several flights of stairs, e - climb one flight of stairs, g - walk more than a mile, h - walk half a mile, i - walk 100 yards) were considered as indicators of locomotor disability in ordinal logistic regression models. This demonstrated an approach to analysis that moves beyond the simple binary definitions of locomotor disability used in the majority of previous studies (see Chapter 3).

However, the use of these individual items did not produce the interval-level of measurement that was argued for in Chapter 1. The next chapter will take the five locomotor disability-specific items from the PF-10 and use them to create such a measure. The only mathematically proven way to transform ordinal-level item responses, such as those of the PF-10 items, into an interval-level measure is through the use of the Rasch measurement model (Wright & Linacre 1989, Fischer 1995). This chapter therefore presents the background to and theory of this model, before it is applied to the PF-10 items in the following chapter.

9.2 The mathematics of the Rasch model

The Rasch unidimensional measurement model; or more simply, the Rasch model, is a logistic model is named after Georg Rasch, a Danish mathematician who originally developed the methodology (Rasch 1960). Rather than being a mathematical technique to explain a set of data, as is the case with most statistical modelling techniques, the Rasch model is a mathematical specification that data must meet in order to create a unidimensional, interval-level measure (Svetlana Beltyukova personal communication, 25 September 2008). The model creates this measure by combining individual items with binary or ordinal response options. This section describes the rationale behind and the mathematics of the model.

9.2.1 Guttman patterns

In most outcome measures of physical function the items in the measure represent a wide range of task difficulties, for example moving around your home, walking a mile, running a marathon. In general, it is expected that people state that they have difficulty with the 'harder' tasks before they have difficulty with the 'easier' tasks. It is also true that one would generally expect a 'more able' person to have difficulty with fewer items than a 'less able' person. However, it is necessary to check that the empirical orderings are as expected (Wright & Stone 1979).

A Guttman scale is a deterministic model of potential responses to a set of items, based on exactly this principle. Figure 9.1(a) displays the responses of five persons to three hypothetical questionnaire items with binary response options, relating to tasks of varying difficulties. A person scores 1 if they find a task difficult and 0 if they do not find it difficult. For the purposes of the example, it is assumed that there are no missing data. These data can be reordered in terms of the task difficulty and the amount of the trait being assessed that each person demonstrates (Figure 9.1(b)). It can be seen that Task 3 is the most difficult (most people score 1), whilst Task 1 is least difficult (most people score 0). Assuming this hierarchy to be true, the responses of Persons 2, 3, 4 and 5 follow a Guttman pattern, i.e. if the person reports difficulty with an item, then they also report difficulty with all more difficult items and they do not report difficulty with any less difficult items. Person 4 is the most able person (no difficulty with any items), whilst Person 2 is the least able (difficulty with all items). Person 1 does not follow a Guttman pattern because they reported difficulty with the most difficult task (Task 3) and the least difficult task (Task 1), but they did not report difficulty with the task in between (Task 2). Tasks 3 and 2 follow a Guttman pattern, whilst Task 1 does not.

A Guttman pattern relies on a hierarchy within the set of items, and is deterministic. Hence any deviation from the pattern (such as Person 1 in Figure 9.1) will result in a rejection of the pattern in the dataset. In reality, data contain idiosyncrasies and so a pure Guttman pattern is rarely seen (Andrich 1985).

Figure 9.1 An example of a Guttman pattern

a. Responses to hypothetical questionnaire items with binary responses

	Item1	Item 2	Item 3	Total score
Person 1	1	0	1	2
Person 2	1	1	1	3
Person 3	0	1	1	2
Person 4	0	0	0	0
Person 5	0	1	1	2
Total	2	3	4	9

b. Responses to hypothetical questionnaire items with binary responses: ordered by item difficulty and person ability

	Item 3	Item 2	Item 1	Total score
Person 2	1	1	1	3
Person 3	1	1	0	2
Person 5	1	1	0	2
Person 1	1	0	1	2
Person 4	0	0	0	0
Total	4	3	2	9

Under a Guttman pattern, the total score on a set of binary item responses is a sufficient statistic for the distribution of responses (Andrich et al 2003), i.e. knowing a person's total score means knowing how they responded to each item. This is because, under the Guttman pattern, in a hierarchy of *n* items, a score of *s* indicates that the person has difficulty with the first *s* items but not the remaining *n*-*s* items. So, in Figure 9.1, there were four possible scores (0, 1, 2, 3), and under a Guttman pattern, the only way to get each of these scores is to have difficulty with no tasks (score 0), and have difficulty with Task 3 only (score 1), have difficulty with Tasks 3 and 2 only (score 2) or have difficulty with all tasks (score 3). Hence, the total score can be worked back to the actual response pattern (Andrich 1985): the total score is a 'sufficient statistic' for the response pattern.

Box 9.1 illustrates some types of pattern that might be seen in the responses to 12 binary tasks placed in order of difficulty. Clearly under the Pattern 1, the person's ability lies between the

sixth and seventh tasks. Under Pattern 2, the person's ability lies somewhere between the fifth and eighth tasks. In the Patterns 3 and 4, it is not at all clear where the person's ability lies.

Box 9.1 Example response patterns with 12 tasks		
Scoring 1 represents finding a task difficult		
	Easy	*Hard*
Pattern 1	000000111111	
Pattern 2	000001011111	
Pattern 3	101010101010	
Pattern 4	111111000000	

There are many reasons as to why Pattern 1 does not occur in practice. It may be because the items do not go together in a scale, i.e. they may not be unidimensional, and therefore the response to one item may be largely independent of the responses to other items, resulting in a pattern similar to Pattern 3. Alternatively, the items might have approximately equal difficulty, and therefore it is difficult to order the items, again resulting in a pattern resembling Pattern 3 (Andrich et al 2003). In the field of education Pattern 4 may occur if that person has special knowledge of a particular area and so endorses difficult items when they do not endorse easier ones. This scenario seems less likely to occur in the assessment of disability. If the items are indeed in the expected order, then the most likely reason for a failure of the pure Guttman pattern is chance, where, close to their ability, persons are almost equally likely to report difficulty with an item as not. This results in a pattern similar to Pattern 2.

The Rasch model can be seen as using a probabilistic version of a Guttman pattern that allows some deviation from this pattern. Under the Rasch model, tasks are expected to have a hierarchy along a single latent construct, but to be locally independent, i.e. the response to one item does not determine the responses to other items (Marais & Andrich 2008). Hence a pure Guttman pattern might indicate the tasks are too similar, i.e. they are not locally independent, whilst a pattern very dissimilar to Guttman might indicate that the tasks are measuring different constructs, i.e. the scale is not unidimensional.

9.2.2 The dichotomous model

The Rasch model is based on the principle that the probability of a person endorsing an item is a function of two factors, 1) the person's 'ability', i.e. the amount of the latent trait that they possess, and 2) the 'difficulty' of the task, i.e. how much of the latent trait the item requires.

Logically, if person ability and task difficulty are measured on the same scale then people who have more ability than the task is difficult should not report difficulty with the task and those with less ability should report difficulty. This means that, under the assumptions of a Guttman pattern, the difference between the ability of person n, θ_n, and the difficulty of task i, δ_i, is a good way to describe their relationship. However, this is a deterministic approach, and it is more useful to predict the probability of success, rather than to predict absolute success. For example, it is more useful to know that those with more ability are less likely to report difficulty than those with less ability, because in reality, data have idiosyncrasies (Wright & Linacre 1989).

The difference, $\theta_n - \delta_i$, ranges from negative infinity to positive infinity, as this is the range of latent trait. To constrain the measure of probability to the required 0 to 1 range, the logistic function is applied, as in (1), where, p_{ni} is the probability of person n reporting difficulty with task i.

$$p_{ni} = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}} \qquad (1)$$

This is a useful function because it is the only mathematical function for an ogive, i.e. the shape required to describe the relationship between an ordinal-level score and an interval-level measure (Figure 2.2), that also allows θ_n and δ_i to be estimated independently (Wright & Stone 1979). This is important because it is a requirement of fundamental measurement that task or 'item' and person parameters can be estimated separately, so that the item (person) calibrations are independent of the persons (items) that are used to estimate them (Andrich 1985). It is equivalent to saying that you want the length of the piece of string that you measure to be the same, regardless of which ruler you use to measure it (Andrich 1985). It is true that θ_n and δ_i can be estimated separately, because, under a Guttman pattern, they are sufficient statistics for the total scores of persons and items respectively (Andrich 1985).

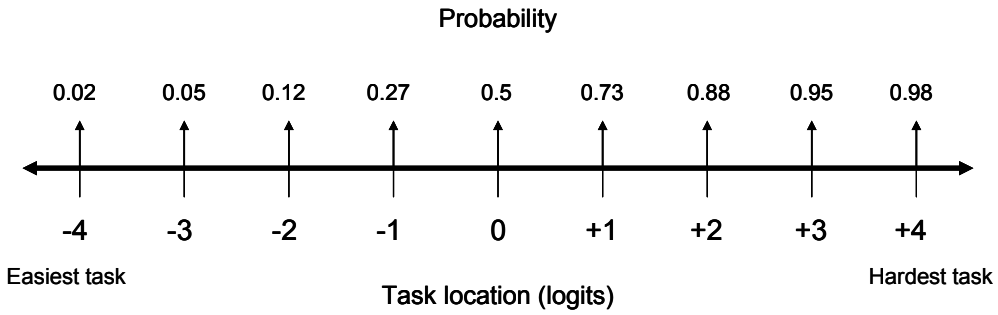
Under the Rasch model, all response patterns are allowable, but those closest to a Guttman pattern are most likely (Andrich 1985). The probability of a Guttman pattern is higher if the item

locations are further apart; when items are of similar difficulty, there is more chance of other patterns occurring (Andrich 1985).

The Rasch model is defined on a logit scale, because of the logistic transformation described above. Persons and items are assigned a value in logits, on the same scale. Assuming higher logit scores mean that a person has more of the trait being measured, or that the task under consideration is more difficult. Ability is the log odds of reporting difficulty with a task that has a difficulty of zero logits. Task difficulty is the log odds of a person with an ability of zero logits not reporting difficulty using the same logit scale. (Wright & Stone 1979).

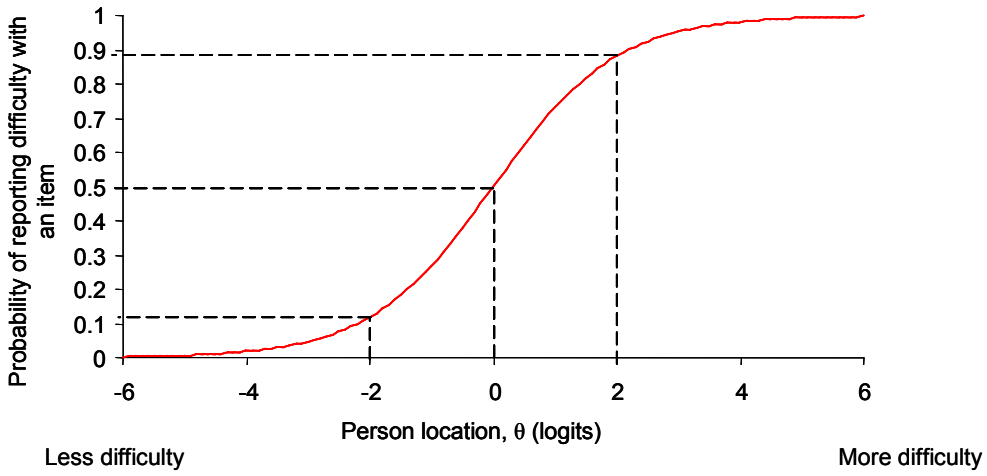
When person and item locations are equal, the probability of the person reporting difficulty on the item is 0.5. The probability of reporting difficulty on other items is based on the distance between the person and item locations, as in Equation 1. Figure 9.2 shows the probability of endorsing items with locations ranging from -4 to +4 logits with a person location of zero logits. These probabilities stem from the difference, in logits, between the person location θ_k and the item location δ_i in the Rasch model (Appendix E). For example, for a person with a location of zero logits, the probability of reporting difficulty on an item that is located at +2 logits is 0.88, whilst the probability of reporting difficulty on an item that is located at -2 logits is 0.12. Hence, people are more likely to report difficulty on items that are 'harder' than they are 'able'. In the example in Figure 9.3, people are more likely to report difficulty on items with locations above zero logits than below.

Figure 9.2 Probability of reporting difficulty on an item at different item locations with a person location of zero logits



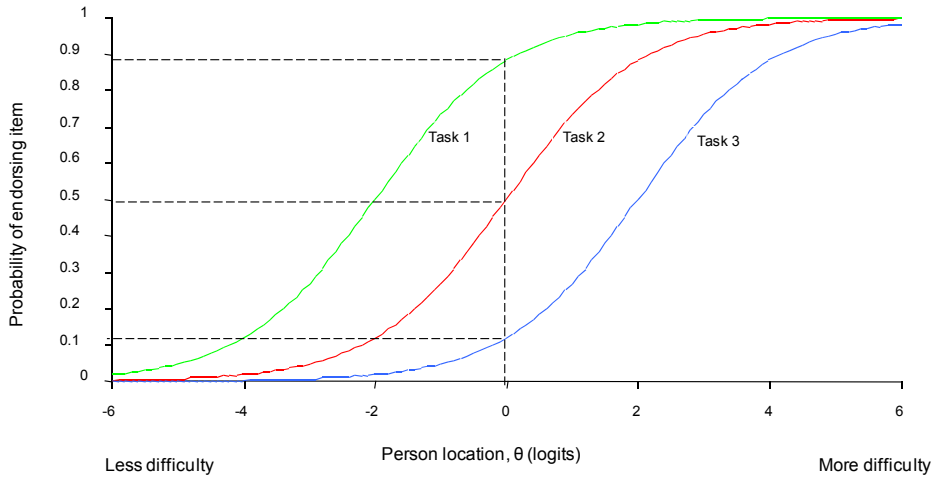
These probabilities can also be presented in terms of the item, on an item characteristic curve, an example of which is given in Figure 9.3. An item characteristic curve plots the position of a person on the latent trait in logits (on the x-axis axis) against the probability of reporting difficulty with an item of specific difficulty. From Figure 9.4, it can be seen that for a person with a location of zero logits, the probability of reporting difficulty with a task of difficulty 0 logits is 0.5. For a person with a location of –2 logits, the probability of difficulty with the same task is 0.12, whilst for a person with a location of +2 logits, the probability of difficulty with the task is 0.88.

Figure 9.3 An example of an item characteristic curve: item location zero logits



The item characteristic curves for several items can be overlaid on the same graph (Figure 9.4). These overlaid curves can give information about task difficulty and person ability. First, consider a person with a location of 0 logits, this person has probability of approximately 0.88 of reporting difficulty with Task 1, a probability of 0.5 of difficulty with Task 2 and a probability of 0.12 of difficulty with Task 3. So it can be seen that Task 1 is the most likely to be found difficult, whilst Task 3 is the least likely to be difficult. Then consider the parallel lines on Figure 9.4 representing Tasks 1, 2 and 3. Any person, regardless of their location on the latent trait, is more likely to find Task 1 difficult than they are Tasks 2 or 3. Everyone is least likely to find Task 3 difficult.

Figure 9.4 Item response curves for hypothetical Tasks 1, 2 and 3



9.2.3 The polytomous model

This discussion of the Rasch model has so far assumed that the items in a scale have dichotomous response options: a person either reports difficulty or they do not. However, the dichotomous Rasch model can be extended to include items with multiple, ordered response options. This can be done in either of two frameworks: the rating scale model (RSM) (Andrich 1978) or the partial credit model (PCM) (Masters 1982). This section will explain both models, their similarities, differences and when they might be used.

Taking an item with *k* ordered response categories there are *k*-1 thresholds between the categories, as described in Chapter 7. So for the PF-10 items with three response options, there are two thresholds: one between "Yes, limited a lot" and "Yes, limited a little", and one between "Yes, limited a little" and "No, not limited at all".

The RSM can deal with categorical item responses by modelling these thresholds between the categories, as well as the person and item locations (2), where, p_{nij} is the probability of person *n* responding in category *j* of item *i.*

$$p_{nij} = \frac{e^{\theta_n - \delta_i - \tau_j}}{1+e^{\theta_n - \delta_i - \tau_j}} \qquad (2)$$

The key property of the RSM is that the distances between thresholds are fixed across items. So, in Figure 9.5(a), where the thresholds are represented by a change in the colour of the bar, the red bar (numbered 1) is the same length for each item and the green bar (numbered 2) is

the same length for each item. However, the red bar and green bars do not need to be the same length as each other. In addition, the overall location of the item, i.e. mean position of the thresholds in logits, does not have to be the same on each item.

The PCM, is similar to the RSM, but without the constraint on the distance between thresholds, i.e. the distances between all thresholds on all items are allowed to vary (Figure 9.5(b)). The PCM is estimated using (3).

$$p_{nij} = \frac{e^{\theta_n - \delta_{ij}}}{1 + e^{\theta_n - \delta_{ij}}} \quad (3)$$

Some authors prefer to use the PCM in most instances, even when the items all have the same response category labels, such as the PF-10, because it allows for differences between thresholds across items (Eyres et al 2005, Pallant et al 2006). Others, however, maintain that the PCM should be reserved for instances in which the number of response categories varies across items, such as the Berkman-Syme Index (Berkman & Syme 1979), or in which the response categories are given different labels for different items (Haley et al 1994), such as the HADS (Zigmond & Snaith 1983). If one does not conform to the principle that the RSM should always be used when response options are the same, then, the respective fits of the RSM and PCM can be tested empirically using a likelihood ratio test, where the RSM is nested within the more complex PCM (Andrich et al 2003).

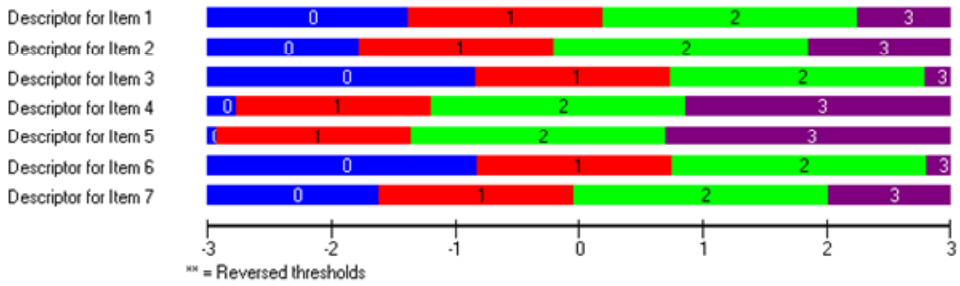
9.3 Practicalities of the Rasch model

9.3.1 Computer packages

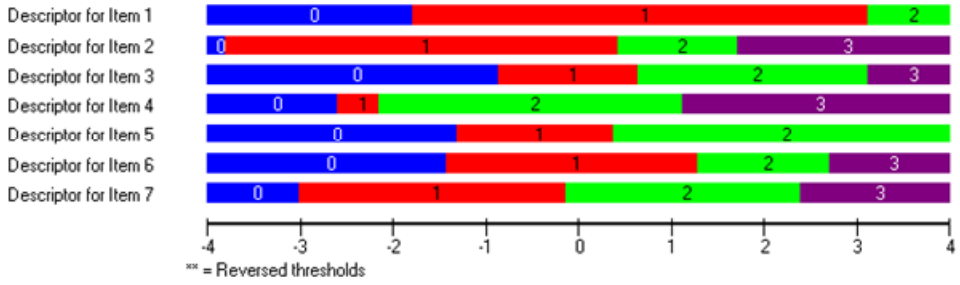
There are several specialist packages available with which to fit data to the Rasch model. These packages include WINSTEPS (Linacre 2007), RUMM2020 (Andrich et al 2003) and ConQuest (Wu et al 1998). All of these packages estimate the Rasch model and all provide tests of the fit of the data to the model. However, the way the model is fitted and the actual tests of fit differ slightly between packages. In this PhD project, the program RUMM2020 was used, and so the description of data fit to the Rasch model and how this can be tested, is based on the RUMM2020 package.

Figure 9.5 Example of threshold maps for the rating scale and partial credit models: four response options per item

(a) Rating Scale Model



(b) Partial Credit Model



9.3.2 Class intervals

Many of the indices of fit in RUMM2020 are based on what are known as class intervals. These are groups of individuals, defined by the rank order of person locations according to the logit score produced by the Rasch estimation procedure (Andrich et al 2003). The number of class intervals is dependent on the sample size, and it is suggested that each class interval should contain around 50 people (Tennant et al 2008).

9.3.3 Sample size

The sample size used in estimating Rasch model parameters is important, to ensure the stability of parameter estimates (Linacre 1994). Linacre (1994) suggests that parameter calibration (for items or person) to within ±0.3 logits across similar samples is as good as can be expected,

whilst Wright and Douglas (1975) suggest that for practical purposes, a calibration within ±0.5 logits is free from bias. For dichotomous items, where approximately half of people endorse each item, Linacre calculated that a sample size of 150 should suffice to have 99% confidence that the parameter estimates are stable to within ±0.5 logits (Linacre 1994). Further work then suggested that, for polytomous items, at least 10 responses in each category are required and hence the sample size will be determined by the number of items in a scale, as well as the distribution of responses to those items (Linacre 2004).

When fitting data to the Rasch model in RUMM2020, there is the added consideration for sample size that the tests of the fit of the data to the Rasch model are based on statistical significance tests. Hence, although a large sample size will give a more precise estimate of the model parameters, it is also likely to result in the rejection of the fit of the data to the Rasch model, even if there are only moderate departures from the specification. It is suggested that an adequate sample size for most purposes is 243 (Tennant et al 2008). In large samples, a random sample of the data could be taken for model fitting purposes, or RUMM2020 allows for an 'effective sample size' to be specified so that the power of the significance tests can be controlled (Andrich et al 2003).

A further consideration in deciding on an adequate sample size is the number of people with 'extreme scores'. That is people giving the lowest or highest response to every item. These people do not provide any information about the relative difficulties of the items; it is only known that they are worse or better off than can be assessed by the items in the scale. These people cannot therefore be used in the estimation of model parameters.

9.3.4 Model estimation procedures

RUMM2020 uses a weighted maximum likelihood-based approach to model estimation (Andrich et al 2003). This procedure iteratively estimates the person and item parameters and can estimate missing item responses. This means that, unlike a summated rating approach, people without complete data can receive a Rasch score.

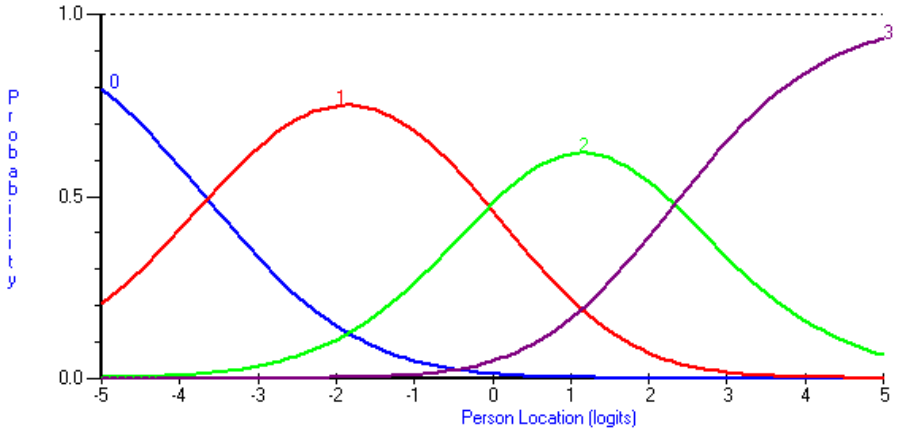
9.4 Testing fit to the Rasch model specifications

9.4.1 Threshold ordering

Under classical test theory, when items have more than two categories, it is usually assumed that respondents have treated these categories in the way that they were intended to be used. This is usually the order in which they were presented in the questionnaire. For example, in the SF-12 there are three items with six response options: “All of the time”, “Most of the time”, “A good bit of the time”, “Some of the time”, “A little of the time”, “None of the time”. However, the exact ordering of the options may not be clear to respondents. Is “A little of the time”, more or less time than “Some of the time”? The Rasch model provides the opportunity to test the assumption that the responses are used in the intended order by providing the empirical ordering of the categories as interpreted by respondents (Tennant & Conaghan 2007).

Figure 9.6 shows a plot of the latent trait on the x-axis against the probability of response on the y-axis for a single item with four response options. Curves 0, 1, 2 and 3 on the plot show the probability distribution of each of the four response categories along the latent trait. It would be expected that every response option must have a point along the latent trait when it is the most likely response, i.e. when reading the plot from left to right, the first threshold (transition from Curve 0 to Curve 1) occurs first, the second threshold (transition from Curve 1 to Curve 2) occurs next, and so on. This is what happens in Figure 9.6.

Figure 9.6 Example of ordered category thresholds in a Rasch model: four response options



However this ordering does not always occur. Where there are many response options, or where there may be potential ambiguity about the relative difficulty of the response options within an item such as in the SF-12 items mentioned above, response categories might not be used consistently across the sample of people completing the item. This results in disordered response categories, as in Figure 9.7(a), where Threshold 2 (transition from Curve 1 to Curve 2) occurs before Threshold 1 (transition from Curve 0 to Curve 1). This means that there is no point on the person score at which the second category is the most likely response option, i.e. no point at which Curve 1 is the highest.

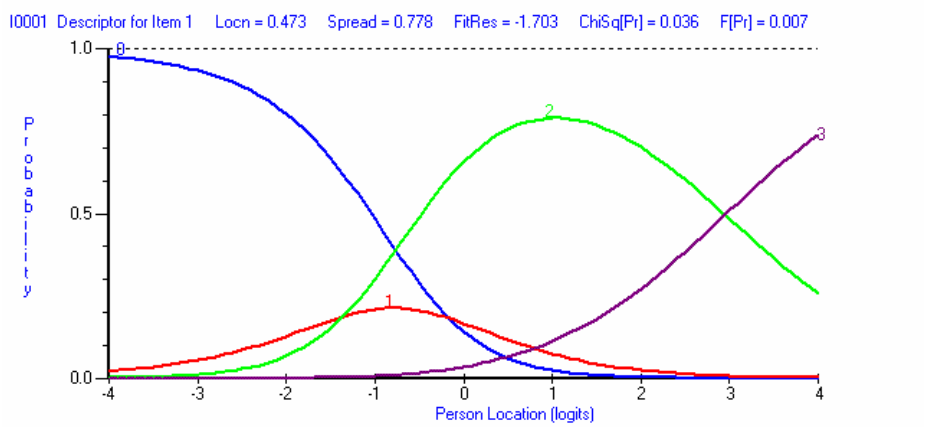
Disordered response categories can result in poor overall fit to the Rasch model specification. This issue can be remedied by collapsing adjacent categories (Tennant et al 2008). So in the example in Figure 9.7(a), the problem can be corrected by combining responses in categories 1 and 2 (Figure 9.7(b)). This results in three rather than four response categories for use in the analysis, but could improve the overall fit of the data to the Rasch model. It is generally advised not to include extreme categories when rescoring an item, i.e. categories 0 and 3 in Figure 9.7(a) should not be collapsed with category 1 or 2, respectively, unless this is unavoidable, because the extreme categories are theoretically assumed to be infinite (Tennant et al 2008).

9.4.2 Unidimensionality

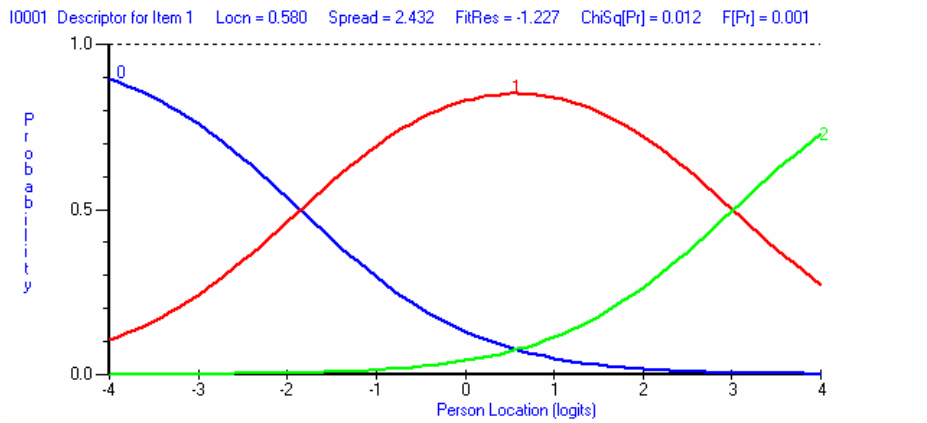
In physical measurement, one would only measure one entity at a time. For example, it does not make sense to compare the weight of two people, based on their body mass index (BMI), which is a combination of their weight and height. Two people with similar BMI might actually have very different weights, because their heights were also different. In the Rasch model, it is assumed that the items form a single construct (Tennant & Conaghan 2007). If this is not the case, then the estimates of the item and person locations can be biased (Smith 2002). That is not to say that several different processes cannot contribute to the way in which items are answered, but the processes must act in the same way across all items (Bejar 1983).

Figure 9.7 Example of disordered category thresholds in a Rasch model

a. Original disordered category thresholds: four outcome categories



b. Rescored item with ordered category thresholds: three outcome categories (original categories 1 and 2 combined)



One way to conceive of this unidimensionality is that once the 'Rasch factor' has been removed there is no further pattern in the data (Tennant & Conaghan 2007). Hence, a principal components analysis of the residuals should show no pattern and all of the extracted components should account for a similar proportion of the variance (Andrich et al 2003).

Within RUMM2020, unidimensionality can also be assessed by means of an independent t-test of person locations. First, a principal components analysis of the residuals is carried out to ascertain the two most different groups of items within the item set, i.e. those that load positively onto the first factor and those that load negatively. Two person locations are derived for each

individual, one based on the positively loading items and one based on the negatively loading items. An independent t-test (because the two sets of items are independent) is then used to determine whether the two person locations are significantly different. It is suggested that for a scale to be considered unidimensional, no more than 5% of people should have person locations from the two item sets that are significantly different at the 5% level (Tennant & Conaghan 2007).

Tennant and Conaghan (2007) suggest that these t-tests for unidimensionality are for strict rather than essential, unidimensionality. Hence, when the assumption of unidimensionality does not hold, further investigations should be carried out to identify possible multiple dimensions within the item set.

Unidimensionality can also be assessed via the correlation between the item residuals. Strong negative correlation can be a sign of multidimensionality in the scale. However, these correlations generally tend towards the negative and when there are only two items it is not unusual for their residuals to show almost perfect negative correlation (David Andrich personal communication, 14 September 2008).

9.4.3 Local response dependency

Local response dependency is an important issue in the fit of data to the Rasch model. It occurs when the response to one item is determined by the response to another item in the scale (Tennant & Conaghan 2007). For example, with two items, the first regarding carrying one bag and the second regarding carrying many bags, if a person responds that they can carry many bags, then, logically, they must also respond that they can carry one bag.

It should usually be possible to anticipate where local response dependency will occur, but it is also important to consider empirical evidence. The Rasch model assumes the statistical independence of the responses to items (Marais & Andrich 2008, Tennant & Conaghan 2007), so too strong a relationship between the responses to items can bias the parameter estimates in the model.

It should be noted that local response dependency does not occur because items have similar locations. Items with similar locations result in more precise estimates of person location in the area of the latent trait around the items (Andrich et al 2003).

Like unidimensionality, local response dependency can be assessed using the residual correlations between items after the 'Rasch factor' has been extracted, because there should not be any remaining patterns (Tennant & Conaghan 2007). Strong positive residual correlations between items are indicative of local response dependency (Tennant & Conaghan 2007). The interpretation of what is a strong correlation varies, but Tennant et al (2008) suggest that it is a correlation with an absolute value >0.3, whilst Andrich et al (2003) suggest the more qualitative definition of a correlation that is "*noticeably higher than the correlations of residuals between most pairs of items*".

Local response dependency can be accounted for in the Rasch model by forming 'subtests' (Tennant & Conaghan 2007) where the related items are combined into a single item. Hence, in the bag carrying example above, the subtest would effectively be an item with theoretical response categories being the number of bags that the person can carry. Subtests can be created to combine two or more items, and are usually generated within software packages such as RUMM2020.

9.4.4 Item fit

There is no single statistic that can adequately assess the fit of items to the Rasch model. Instead, a series of statistics are needed that assess item fit from "*different angles*" (Tennant & Conaghan 2007).

Most fit statistics are based on the difference between the expected response under the Rasch model and that observed in the dataset, i.e. the residuals (Tennant & Conaghan 2007). Two main methods of assessing item fit, based on residuals, are the item fit residual and the chi-square test.

The item fit residual is calculated as the difference between the observed and expected responses to an item, summed over all persons (Tennant & Conaghan 2007). The value is then standardised to a z-score, i.e. an observation from the Standard Normal Distribution, mean of zero, standard deviation (SD) of 1. Therefore, values outside the range -2.5 to +2.5 indicate that the data are not a good fit to the model at the 99% confidence level. Large negative residuals generally indicate item redundancy, i.e. it may be possible to remove the item from the scale without much

loss of information. Large positive residuals are indicative of under discrimination, i.e. the item does not distinguish between people with different levels of the latent trait (Tennant et al 2008).

The chi-square statistics are calculated in each class interval for each item, based on the difference between what is expected for each individual in the sample, under the Rasch model specification, and what is observed in the data. These chi-square statistics are then summed across the class intervals to get an overall chi-square statistic for the item. This is tested against a chi-square c-1 distribution, where c is the number of class intervals (Tennant & Conaghan 2007, Andrich et al 2003).

An F-statistic can also be calculated; this is very similar to the chi-square statistic except that rather than being based on the observed and expected scores of individual persons, it is based on the observed and expected scores on average in the class intervals (Tennant et al 2008). The F-statistic is compared to an F-distribution with (c-1, n-c-1) degrees of freedom where n is the number of persons. The F-statistic is a more sensitive indicator of the fit of an item to the Rasch model than the chi-square statistic (Andrich et al 2003).

It is usual to use a 5% level of significance for both the chi-square and F-tests, and to make a Bonferroni adjustment for the number of items in the scale (Tennant & Conaghan 2007). For example with 10 items in a scale, significant misfit would be suggested by p-values from the chi-square test of less than 0.005, i.e. 0.05/10.

If all of the tests of item fit point to either good or poor fit of the item to the Rasch model specification, then the fit of the item is easily assessed. However, this is not always the case. In particular, the chi-square statistic can be influenced by a large chi-square value in a single class interval, especially if that class interval is at the extremes of the person distribution, or if the probability of endorsing an item is close to 0 or 1 (Andrich et al 2003). Conflict between the item residual and item chi-square test is most likely to occur when the targeting of the items to persons is not very good, i.e. the respondents, on average, have a higher or lower level of ability than is measured by the scale (see below) (Andrich et al 2003). It is suggested that all statistics relating to fit are examined relatively rather than absolutely and that the final judgement about the fit of an item should also consider the intended use of the scale (Andrich et al 2003). For example, will the scale be used to make individual or group level comparisons?

9.4.5 Person fit

It is important to examine the fit of persons was well as items to the Rasch model specification (Tennant & Conaghan 2007). This can be done using standardised person residuals, in a similar way to their use in assessing item fit. As for item fit, the person fit standardised residuals are expected to be Normally distributed with mean of 0 and a standard deviation of 1, with residuals outside the range -2.5 to +2.5, considered to show misfit (Tennant et al 2008, Andrich et al 2003). The residual statistic is in fact assessing the degree to which an individual has produced a Guttman pattern in their responses: the higher (more positive) the person residual statistic, the further the person is from giving a Guttman response pattern (Andrich et al 2003). In the case of a misfitting person, a large residual indicates that the response pattern did not fit that expected by the Rasch model. It is suggested that it is worth further investigation into such people as they may be able to inform future developments of the scale. For example, they may have a cognitive impairment or comorbid disease that makes them respond differently from the sample as a whole (Tennant & Congahan 2007).

When a sample contains individuals who misfit to the Rasch model specification, it may be appropriate to remove these people from the development stage of the tool so that that data conform to the Rasch model. This approach may however bring into question the generalisability of the measure to populations with such impairments or morbidities (Tennant & Conaghan 2007). There are no ‘rules’ as to when one should remove individuals from a sample, and so it is up to the scale developer to decide whether the misfit is relevant, whether it affects the overall fit of the model and whether removing people will reduce the usefulness of the scale in other populations.

9.4.6 Overall fit

As well as assessing the fit of the individual persons and items to the Rasch model specification, RUMM2020 provides overall fit statistics for person and items in the form of the mean item fit residual and the mean person fit residual. If the data are a good fit to the Rasch model, the residuals should follow an approximate standard Normal distribution. It is therefore expected that the distribution of the fit residual will have a mean of approximately 0 and a standard deviation of 1 (Andrich et al 2003).

The overall fit of the data to the Rasch model can also be tested via a chi-square test. This is calculated by summing the chi-square values for the individual items to generate an overall chi-square value known as the item-trait interaction statistic (Pallant et al 2006). A significant item-trait interaction statistic (chi-square test on $m(c\text{-}1)$ degrees of freedom, where m is the number of items and c is the number of class intervals), after Bonferroni adjustment for the number of items, indicates that the items are not invariant across the trait, i.e. the hierarchical ordering of the items is not the same at all levels of the latent trait and so the data are a poor fit to the Rasch model (Tennant & Conaghan 2007).

9.4.7 Differential item functioning

It is important to ensure that items behave in the same way in different groups of people (Tennant & Conaghan 2007). For example, are males and females with the same amount of the latent trait equally likely to endorse an item? When different groups respond to an item differently, despite the same level of the latent trait, this is known as differential item functioning (DIF). DIF can take two forms: uniform and non-uniform.

Uniform DIF occurs when one group responds systematically differently to another group along the latent trait. For example, Figure 9.8(a) shows the response of males and females to a polytomous item. The x-axis shows the latent trait in logits and the y-axis shows the expected response to the individual item, which in this case has four possible options (0, 1, 2, 3). At each point along the latent trait, females have a higher expected response than males and the difference between the expected male and female scores is approximately equal along the trait.

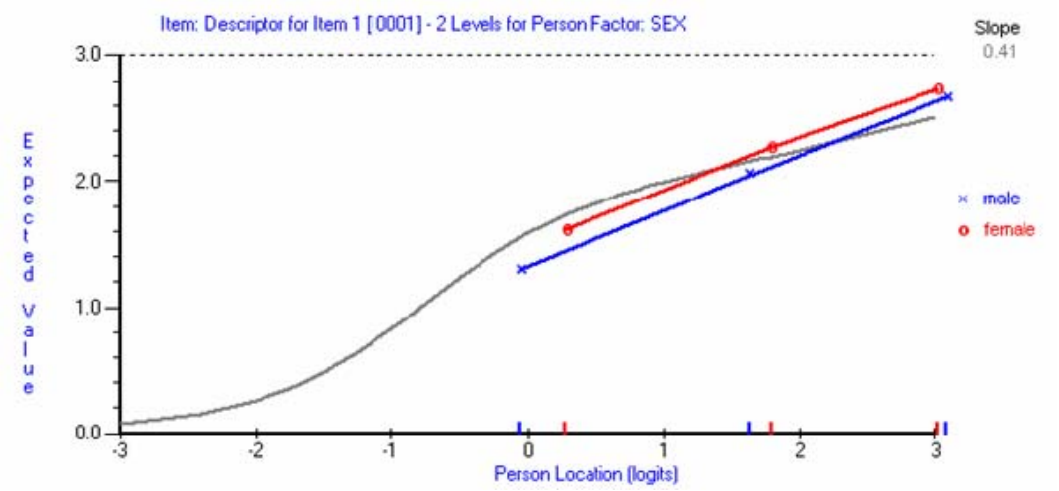
Non-uniform DIF occurs when the response of one group is different from the response of another group and this difference changes, i.e. is non-uniform, along the latent trait. Figure 9.8(b) shows the response of males and females to a particular item. At the lower end of the latent trait (below 1 logit), females have a higher expected response than males, but at the higher end of the latent trait (above 2 logits), males have a higher expected response than females.

Clearly it is possible that different groups can display more of the latent trait than others. For example, females may on average be more depressed than males, but when taking males and

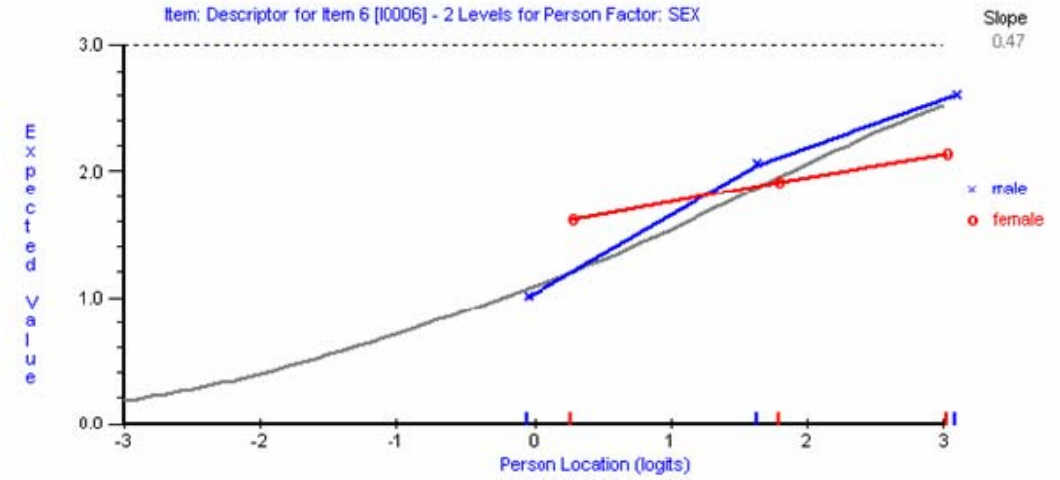
females with the same amount of depression, say, it is expected that they respond to an item in the same way (Tennant & Congahan 2007).

Figure 9.8 Example of differential item functioning in a polytomous item with four response categories

a. Uniform differential item functioning



b. Non-uniform differential item functioning



Each individual item can be tested for DIF within the RUMM2020 package by means of an F-test, from an analysis of variance on the standardised residuals. A p-value less than 0.05 for the main effect, for example gender, is an indication of uniform DIF. A p-value of less than 0.05 for the

interaction term, for example between gender and the class intervals, is indicative of non-uniform DIF. When developing a new scale it is usual not to apply a Bonferroni adjustment to the results of the F-test, in order to be more conservative. However, the Bonferroni adjustment is often made in validation studies (Alan Tennant personal communication, 13 March 2008).

DIF can be displayed for two main reasons. First, it can be true DIF. For example, in a measure of cognitive function, the question "During what years did the First World War take place?" might be answered correctly more often by males than by females, despite the same level of cognitive functioning, because males generally take more of an interest in this topic. Second, DIF can be artefactual. This occurs when there is real DIF in another item or items and, because of the way DIF is assessed, it needs to be 'balanced out' somewhere else. This type of DIF will disappear if the true DIF is somehow removed from the scale.

When an item displays DIF, it is a sign that the item is breaching the assumption of unidimensionality in the scale (Pallant et al 2006). It is possible, when faced with uniform DIF, to 'split the item for DIF' (Tennant et al 2008). This effectively means treating the item separately for each group in the factor displaying DIF. So, in the example in Figure 9.8(a), in the overall scale score, there would be one 'item score' for males and one 'item score' for females instead of the one original item score for everyone. There is little guidance in the literature on how to deal with non-uniform DIF, and some authors suggest that any item with DIF, be it uniform or non-uniform, should be removed from an item set because it displays bias (Pallant et al 2006). It is also worth considering whether the DIF is important (Svetlana Beltyukova personal communication, 25 September 2008), as well as statistically significant – it may be that the DIF is inconsequential and does not affect the overall fit of the data to the Rasch model, in these circumstances it may be appropriate to leave the item with DIF in the scale, as it is.

It is suggested that as a minimum, DIF should be investigated with respect to age-groups and gender (Tennant & Conaghan 2007). However, in certain instances in may be necessary to test for DIF by other factors, such as culture or patient group. The decision about which factors are used to investigate DIF depends on the use of the tool and the reason for the analysis (Svetlana Beltyukova personal communication, 25 September 2008). In the case of this thesis, analysis for DIF will be carried out in the context of the development of a new measure. Therefore, DIF is of

interest in relation to item bias, i.e. to examine whether different groups in the population respond differently to items, and will only be investigated by age and gender.

9.4.8 Targeting

The targeting of the scale refers to the relative locations of the respondents and the items: are the majority of items located in the same area of the scale as the persons? Targeting is important because a well-targeted scale allows better measurement of the attribute in question than a poorly targeted scale. Targeting is also related to floor and ceiling effects and the responsiveness of a measure over time (see Chapter 5).

In RUMM2020, the mean item location is constrained to be 0 logits (Andrich et al 2003). Hence, a negative mean person location suggests that the sample has less of the latent trait than can be measured by the items, whilst a positive person location suggests that the sample has more of the trait than the scale measures.

The relative position of persons and items can be summarised by the Person Separation Index (PSI). The PSI ranges from zero to one, with values closer to one indicating that the scale has more ability to distinguish between persons. It is generally suggested that the PSI be interpreted in a similar way to Cronbach's alpha, with a value of 0.7 necessary for comparison between groups (Tennant et al 2008).

9.5 Implementation for the locomotor disability measure in other samples

When items have been confirmed as fitting to the Rasch model specification in a specific dataset, it would often be desirable to create the logit score in other datasets. This section outlines two possible approaches to this: the Rasch model or a score conversion table.

9.5.1 Potential methods

9.5.1.1 Use of the Rasch model

New data can be fitted to the Rasch model, allowing a Rasch transformed interval-level measure to be computed for persons in these data. These person estimates would have a slightly

different range in each of the datasets fitted to the Rasch model, as slightly different parameter estimates would be calculated in each dataset. This may be seen as problematic, especially when the samples are very similar, for example random samples from a single, larger population. However, this is the method that will result in the most accurate Rasch person location estimates for an individual sample.

Furthermore, this method of generating Rasch scores in new data has the advantage that scores can be computed for people who have not completed all items. This is potentially important, depending on levels of missing data. Also, using the Rasch model in the new dataset, it is possible to assess the fit of the data to the model. However, with epidemiological datasets likely to be large, and some tests of fit being based on statistical significance, it is likely that data would appear to misfit the Rasch model, due to the large sample size. This could be overcome by specifying an effective sample size for the analysis (Andrich et al 2003).

If the dataset is longitudinal, with the outcome of interest available at more than one time point, then using the Rasch model to estimate the person locations over time can ensure that there is no DIF across time points and that the 'ruler' does not change over time (Wright 2003).

The major barrier to the use of Rasch analysis to create the locomotor disability score in other populations is the requirement that the person wanting to create the score have access to appropriate software, for example RUMM2020, WinSteps or ConQuest. Recently, it has become possible to estimate a Rasch model in more general purpose statistical packages such as R (Li 2006) and MPlus (Muthén & Muthén 1998-2010), but this still requires a great deal of knowledge of both the model and the software, and in most circumstances would probably not be practical.

9.5.1.2 Use of a conversion table

An alternative to using the Rasch model to create the score in a new dataset is a conversion table. This is a simple way in which the raw sum scores can be converted to an interval-level measure, and such a table is provided by the RUMM2020 package. The disadvantage of this approach is that it cannot give a score for those people who have any missing item responses.

Also, the conversion table is based on the sample in which the measure was developed and therefore it may not be appropriate to apply the conversion table. This is particularly relevant if the sample was not representative or the sample was small, resulting in unstable parameter estimates.

Furthermore, the estimates of person location for individuals in new datasets will not be as accurate as if the Rasch model had been used to create them. Most importantly, the score may not then be the assumed interval-level measure.

9.5.2 Summary

There is little evidence around the best way to derive an interval-level Rasch measure in a new sample, but arguments point towards the use of the Rasch model in new datasets to create dataset-specific Rasch-transformed interval-level measures, despite the potential practical difficulties of this approach. This method is also particularly useful in longitudinal datasets where the score can be 'anchored' over time.

9.6 Practical derivation of an interval-level measure using the Rasch model in longitudinal datasets

As suggested above, using the Rasch model, it is possible to create a Rasch transformed measure across time points in a longitudinal study that does not itself 'change over time' (Wright 2003). This section describes how this process can be performed.

First, the data are 'stacked', i.e. made into a long format, so that each person had a row of data for each time point in which they took part in the study. A variable is created to represent the time points (Figure 9.9). This enables the investigation of DIF by time point. In this format, each item appears in the dataset once for each person at each time point. In the example in Figure 9.9, there are three time points and so each person has responded to each item up to three times. The use of the Rasch model to create the measure means that a missing item at a time point does not preclude the estimation of a score at that time point. The Rasch score is then calculated for each time point, dependent on the other time points, thus ensuring that the 'ruler' does not change over time (Wright 2003).

Figure 9.9 Example of stacked data for estimation of an interval-level score over three time points using Rasch analysis

ID	Time	Item 1 score	Item 2 score	Item 3 score	Item 4 score
1	1	3	4	4	2
2	1	5	5	5	4
3	1	4	2	1	5
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1	2	4	4	4	2
2	2	0	0	3	1
3	2	3	3	0	3
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1	3	4	2	0	1
2	3	4	5	2	4
3	3	4	1	4	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

9.7 Summary and discussion

This chapter has described the mathematics behind the Rasch model and its practical implementation in the RUMM2020 computer package (Andrich et al 2003). It has also described the process of testing the fit of data to the model specification in this package.

This process of model checking can be difficult, as there are at best only guidelines on the values that statistics take when the data are a good fit. At worst, there is little suggestion as to how to conduct these analyses, and one is left to make an educated guess. This is partly the nature of the Rasch model, in that the targeting of the items and persons, the presence of DIF and many other factors influence the fit of the model and the behaviour of fit statistics (Andrich et al 2003). It is also due to the developing nature of the field of Rasch analysis. For example, there have been several studies (Linacre 1994, Wright & Tennant 1996, Smith et al 2008) into the minimum sample size required to carry out a Rasch analysis, but there is little consensus as to what this is in practice.

The Rasch model has been well known in educational research for many years, but until recently, uptake of the model in health research has been more limited (Bond 2008). This is unfortunate given the advantages of this approach. First, and most important for this thesis, is the

potential of the Rasch model to create an interval-level measure, opposed to a long ordinal scale. This model has been proven to be the only way to transform ordinal-level item responses into an interval-level measure (Fischer 1995). Second, the estimates of both person and item locations along the underlying continuum are essentially sample-free (Karabatsos 2004) and can be estimated independently of each other. This is a key property of the Rasch model as it means that the amount of the underlying trait measured does not depend on the ruler used to measure it. Third, in general, the Rasch model can help to avoid the issue of missing data, as scale scores can be computed for those who have completed only some of the items in a scale (Tennant & Conaghan 2007).

Despite some of the potential difficulties in implementing the Rasch model, it is the only method by which to produce interval-level measurement from ordinal items responses, such as response to the PF-10 locomotor disability-specific items (Fischer 1995). For this reason, the next chapter uses the Rasch model methodology to create a scoring mechanism to derive measure of locomotor disability from these five items in the NorStOP pilot dataset. The generalisability of the scoring mechanism to other datasets in considered in Chapter 11, whilst Chapter 12 considers the psychometric properties of the measure against the criteria laid out in Chapter 5.

10 Derivation of an interval-level measure of locomotor disability using items from the PF-10

10.1 Introduction

Throughout this thesis a case has been made for an interval-level measure of locomotor disability. Chapter 6 identified items from the PF-10 that related specifically to locomotor disability and in themselves have good psychometric properties including repeatability over time, validity and feasibility in a research setting. The original scoring mechanism for the PF-10 - summated ratings - was judged to be inappropriate for the formation of an interval-level measure (see Chapter 6), instead forming only a long ordinal-level scale.

Chapter 9 presented the Rasch unidimensional measurement model, which is the only mathematically proven method for the transformation of ordinal-level item responses, such as those from the PF-10 items, into an interval-level measure. This chapter will use the Rasch methodology described in Chapter 9 to devise a measure of locomotor disability from the responses to the PF-10 locomotor disability-specific items in the NorStOP pilot study dataset. The next chapter will test the generalisability of the scoring mechanism for the measure to other samples before its psychometric properties are tested in Chapter 12.

10.2 Initial measure development

The five locomotor disability-specific items from the PF-10 clearly form two separate groups: walking and stair-climbing. Within these groups, there is the obvious potential for response dependency. Although there was no direct evidence of local response dependency in the form of large positive residual correlations, the data were a poor fit to the Rasch model when responses to the five individual items from the NorStOP pilot study were entered into RUMM2020 (Appendix F).

Subtests were used within the RUMM2020 package in an attempt to overcome this problem of poor fit. However, this was not successful, with poor fit to the Rasch model remaining (Appendix 10.2). It was therefore decided to create subtests by studying the relative responses to the

individual items, as suggested by Andrich (1985). In order to distinguish between the subtests created by RUMM2020 and those created on an empirical basis in this chapter, these manually created subtests will be referred to as '***super-items***'.

The remainder of this section on the methods of deriving a measure of locomotor disability from the five locomotor disability-specific items of the PF-10 follows on from this point in the analysis: the five items from the PF-10 cannot be used to form a measure in their raw state.

10.3 Development of super-items

Had the items to be used in this new measure of locomotor disability been dichotomous, i.e. respondents either did or did not report difficulty on each item, then it would be straightforward to expect that for example, those people who could climb several flights of stairs should also be able to climb one flight. In that situation, two super-items, could have been formed, whose categories would have been:

Stair-climbing

1 - cannot climb one flight of stairs;

2 - can climb one flight of stairs but not several;

3 - can climb several flights of stairs;

Walking

1 - cannot walk 100 yards;

2 - can walk 100 yards, but not half a mile;

3 - can walk half a mile, but not more than a mile;

4 - can walk more than a mile.

However, the PF-10 items under consideration have polytomous response options and so, although the average ordering of the items might be obvious, the ordering of the thresholds is far from clear. There are however, some logical rules that could be followed in order to establish such a threshold ordering for each super-item.

- Within an item, Threshold 2 (a little limitation → a lot of limitation) occurs at a higher level of locomotor disability than Threshold 1 (no limitation → a little limitation);
- Threshold 1 for climbing one flight of stairs occurs at a higher level of locomotor disability than Threshold 1 for climbing several flights of stairs;
- Threshold 2 for climbing one flight of stairs occurs at a higher level of locomotor disability than Threshold 2 for climbing several flights of stairs;
- Threshold 1 for walking 100 yards occurs at a higher level of locomotor disability than Threshold 1 for walking half a mile or more than a mile;
- Threshold 2 for walking 100 yards occurs at a higher level of locomotor disability than Threshold 2 for walking half a mile or more than a mile;
- Threshold 1 for walking half a mile occurs at a higher level of locomotor disability than Threshold 1 for walking more than a mile.
- Threshold 2 for walking half a mile occurs at a higher level of locomotor disability than Threshold 2 for walking more than a mile.

These seven logical rules give rise to two possible, logical orderings of the stair-climbing item thresholds and four possible, logical orderings for the walking item thresholds (threshold numbers shown in brackets).

Stair-climbing

a. Several flights (1), Several flights (2), One flight (1), One flight (2);
b. Several flights (1), One flight (1), Several flights (2), One flight (2).

Walking

a. More than a mile (1), More than a mile (2), Half a mile (1), Half a mile (2), 100 yards (1), 100 yards (2);
b. More than a mile (1), Half a mile (1), More than a mile (2), Half a mile (2), 100 yards (1), 100 yards (2);
c. More than a mile (1), More than a mile (2), Half a mile (1), 100 yards (1), Half a mile (2), 100 yards (2).

d. More than a mile (1), Half a mile (1), More than a mile (2), 100 yards (1), Half a mile (2), 100 yards (2);

Table 10.1 shows how often the possible threshold orderings for both the stair-climbing and walking items produced a Guttman pattern in the NorStOP pilot dataset. All threshold orderings were well used, but it was necessary to choose only one for stair-climbing and one for walking, because they were to form the thresholds for the super-items. The most commonly seen orderings, overall and across age and gender groups, were Ordering b for stair-climbing and Ordering d for walking. It was therefore decided to adopt these threshold orderings to create the two super-items for use in the Rasch analysis to generate the locomotor disability measure.

Table 10.1 Ordering of thresholds and their empirical usage in the NorStOP pilot data set, n=358

Ordering[a]		People displaying an allowed response pattern under threshold ordering, n (%)				
		Overall	Gender		Age-group (years)	
			Male	Female	50 to 65	65 and over
Stair-climbing	a	322 (88.7)	152 (87.7)	170 (89.5)	172 (92.0)	150 (85.2)
	b	353 (97.3)	167 (96.5)	186 (97.9)	183 (97.9)	170 (96.6)
Walking	a	292 (80.4)	142 (82.6)	150 (78.5)	156 (83.0)	136 (77.7)
	b	314 (86.5)	156 (90.7)	158 (82.7)	166 (88.3)	148 (84.6)
	c	301 (82.9)	146 (84.9)	155 (81.2)	160 (85.1)	141 (80.6)
	d	323 (89.0)	160 (93.0)	163 (85.3)	170 (90.4)	153 (87.4)

[a]Letters refer to orderings described above for potential super-item thresholds

Table 10.2 shows how the new super-items relate to the categories of the original PF-10 stair-climbing and walking items, assuming that responses conformed to a Guttman pattern under the ordering suggested above.

Table 10.2 Relationship between the scoring for the super-items and the individual PF-10 item responses

a. Stair-climbing

Stair-climbing		Raw super-item score
One flight	Several flights	
Not limited	Not limited	0
Not limited	Limited a little	1
Limited a little	Limited a little	2
Limited a little	Limited a lot	3
Limited a lot	Limited a lot	4

b. Walking

Walking			Raw super-item score
100 yards	Half a mile	More than a mile	
Not limited	Not limited	Not limited	0
Not limited	Not limited	Limited a little	1
Not limited	Limited a little	Limited a little	2
Not limited	Limited a little	Limited a lot	3
Limited a little	Limited a little	Limited a lot	4
Limited a little	Limited a lot	Limited a lot	5
Limited a lot	Limited a lot	Limited a lot	6

10.4 Fit of the super-items to the Rasch model

The two super-items were entered into RUMM2020 and the parameters of the Rasch model were estimated. The Partial Credit Model was chosen, as each super-item has a different number of response categories. In order to create a measure where a higher score was indicative of a higher level of locomotor disability, the scoring of the individual PF-10 items was reversed from that used in the PF-10 sum score, i.e. “yes, limited a lot” was scored 3, “yes, limited a little” was scored 2 and “no, not limited at all” was scored 1.

These analyses were conducted in the NorStOP pilot dataset. The values of fit indices described in Chapter 9 are used in this section to differentiate good and poor fit of the super-items to the Rasch model specification.

10.4.1 Threshold ordering

Initially, both super-items displayed disordered thresholds (Figure 10.1). As discussed in Chapter 9, disordered thresholds occur when individual response categories cannot be easily distinguished. When dealing with super-items, this does not apply and the disordering should not be a concern, as the super-items themselves were not completed by the respondents, rather they were generated by the researcher (Alan Tennant personal communication, 13 March 2008). However, the data did not fit the Rasch model well (Appendix F). Combining categories simplified the data somewhat and made the data a better fit to the Rasch model.

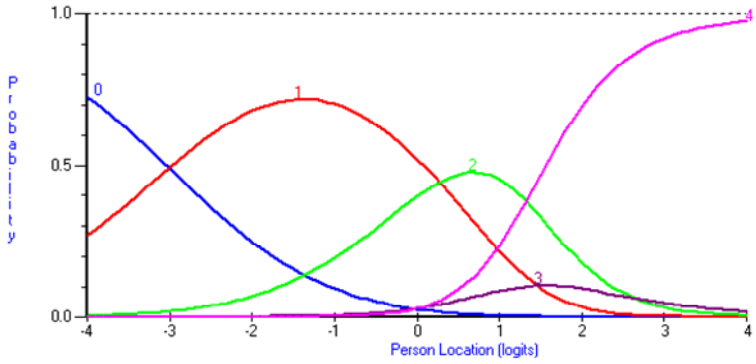
Figure 10.1(a) shows that, for the stair-climbing super-item, the response pattern 'limited a little in climbing one flight of stairs and limited a lot in climbing several flights' (raw score=3 in Table 10.2(a) and Figure 10.1(a)) was not the most likely response pattern at any point along the latent trait. This was remedied by combining the third and fourth response categories for this super-item. The fourth category was not combined with the fifth category, because this is at an extreme of the item distribution.

For the walking item, the response pattern 'not limited in walking 100 yards or half a mile, limited a little in walking a mile' (raw score=1 in Table 10.2(b) and Figure 10.1(b)) was never the most likely response pattern at any point along the latent trait (Figure 10.1(b)) and so, it was combined with the next category along the latent trait (raw score=2 in Table 10.2(b) and Figure 10.1(b)). It was not combined with the category with a raw score of zero, as this was at the extreme of the item distribution.

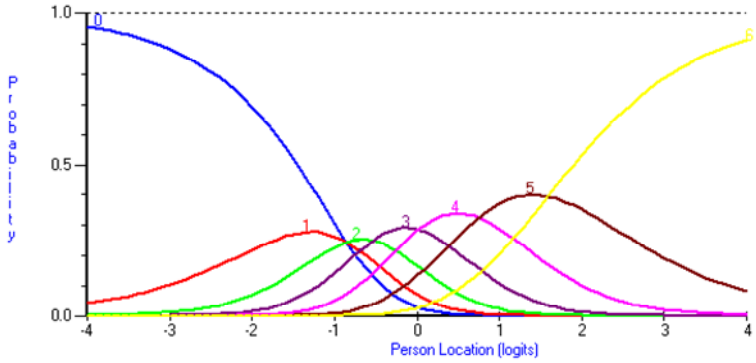
This reordering solved the problem of disordering in the stair-climbing super-item, but resulted in the disordering of two other thresholds in the walking super-item (Figure 10.2). Hence, the new second and third categories (labelled 1 and 2 in Figure 10.2) and the new fourth and fifth categories (labelled 3 and 4 in Figure 10.2) of this item were combined to create a super-item with four categories. The stair-climbing super-item remained unchanged at this step (Figure 10.3(a)).

Figure 10.1 Category probability curves for the stair-climbing and walking super-items

(a) Stair-climbing



(b) Walking



Values assigned to the curves correspond to the raw super-item scores in Table 10.3

Figure 10.2 Category probability curves for the walking super-item having combined the first and second categories of the original walking super-item

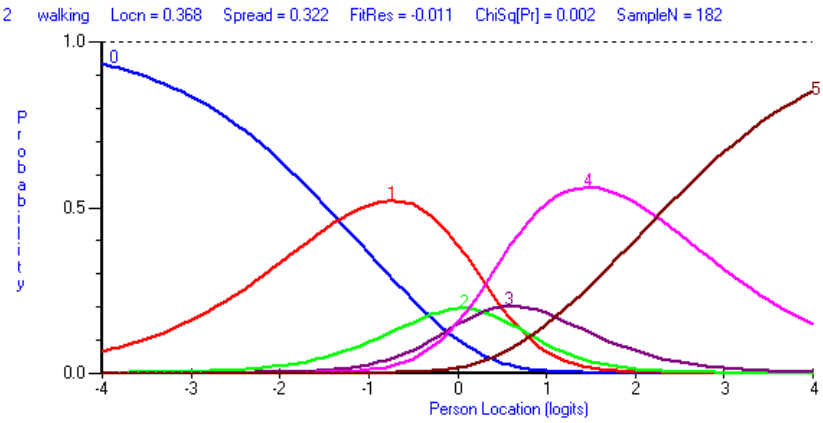
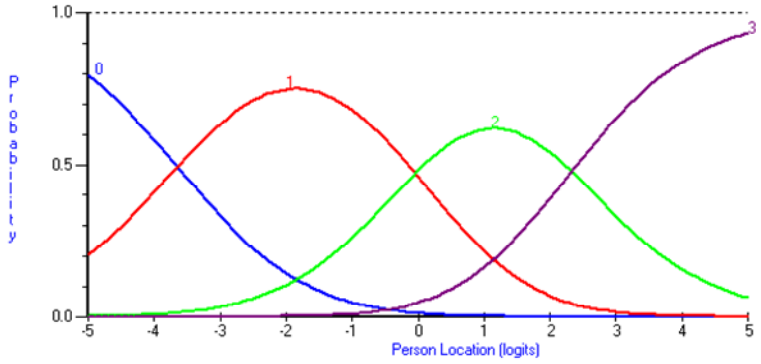


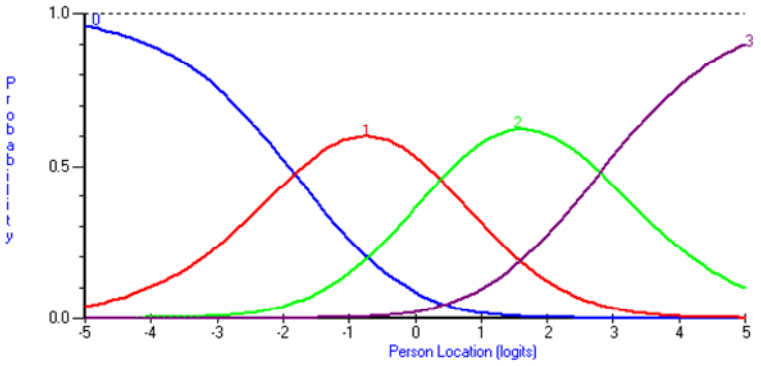
Figure 10.3 shows the category probability curves for the stair-climbing and walking super-items after the second round of rescoring. Under this scoring mechanism, both super-items displayed ordered category thresholds and were essentially simpler than before. Despite both items having four categories, a partial credit model was used, as the same labels could not be attached to the response groups of the two items. Table 10.3 shows the transformations from the raw scores to the new scores for (a) the stair-climbing super-item, and (b) the walking super-item.

Figure 10.3 Category probability curves for the rescored stair-climbing and walking super-items

(a) Stair-climbing



(b) Walking



Values assigned to the curves correspond to the rescored super-item scores in Table 10.3

Table 10.3 Relationship of raw and rescored super-items to individual PF-10 item responses

a. Stair-climbing

Stair-climbing		Super-item scores	
One flight	Several flights	Raw[a]	Rescored[b]
Not limited	Not limited	0	0
Not limited	Limited a little	1	1
Limited a little	Limited a little	2	2
Limited a little	Limited a lot	3	2
Limited a lot	Limited a lot	4	3

b. Walking

Walking			Super-item scores	
100 yards	Half a mile	More than a mile	Raw[a]	Rescored[b]
Not limited	Not limited	Not limited	0	0
Not limited	Not limited	Limited a little	1	1
Not limited	Limited a little	Limited a little	2	1
Not limited	Limited a little	Limited a lot	3	1
Limited a little	Limited a little	Limited a lot	4	2
Limited a little	Limited a lot	Limited a lot	5	2
Limited a lot	Limited a lot	Limited a lot	6	3

[a]As entered into RUMM2020; [b]After collapsing adjacent categories

10.4.2 Unidimensionality

Independent t-tests to assess potential lack of unidimensionality in the super-items showed that 0% (95% CI -3.9%, 3.9%) of persons had significantly different scores at the 5% level based on the stair-climbing super-item alone compared to the walking super-item alone. Hence, there was no evidence against the unidimensionality of the scale.

10.4.3 Response dependency

Residual correlation between the two super-items was -1.0. This is to be expected with so few items and is not necessarily evidence against unidimensionality, especially in light of the results of the independent t-tests described above.

10.4.4 Item fit

The mean item fit residual was 0.348 (SD 0.064). This standard deviation is low compared to the expected value of 1.0 under the Rasch model.

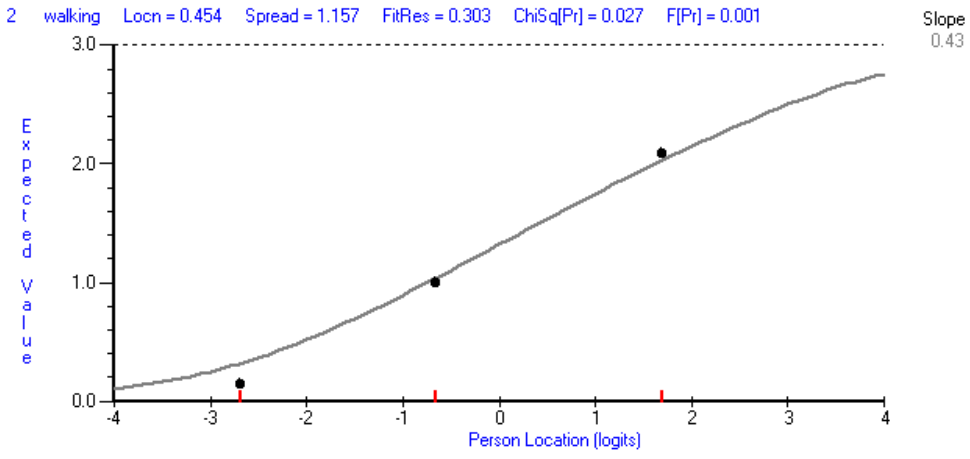
There was no misfit on the stair-climbing super-item as assessed by the standardised residual, the chi-square or the F-test (Table 10.4). On the walking super-item there was no misfit assessed by the standardised residual. However, the chi-square and F-tests showed misfit to the Rasch model specification. Figure 10.4 shows the item characteristic curve for this super-item. The misfit is mainly as a result of those with low levels of locomotor disability (to the left of the plot) reporting slightly less limitation in walking than expected by the model and so should not give cause for concern at this stage.

Table 10.4 Fit of super-items to the Rasch model specification in NorStOP pilot dataset, n=363

Item	Fit residual	Chi-square (df)	p-value	F (df1, df2)	p-value
Stair-climbing	0.393	1.442 (2)	0.4862	1.617 (2, 145)	0.2020
Walking	0.303	7.247 (2)	0.0267	6.925 (2, 151)	0.0013

df - degrees of freedom

Figure 10.4 Item characteristic curve for walking super-item in NorStOP pilot dataset



10.4.5 Person fit

The mean person fit residual was -0.524 (SD 0.732). This mean value was low, compared to the expected value of 0. However, person fit residuals ranged from -1.941 to 0.825 indicating that there were no persons whose responses did not fit the general pattern expected by the Rasch model.

10.4.6 Overall model fit

Under this rescored, super-item model, the item-trait interaction statistic was 8.690 (df=4, p=0.0693). Thus there was no evidence against the assumption of invariance, i.e. the hierarchical ordering of the items remained the same along the latent trait.

10.4.7 Differential item functioning

There was no evidence of DIF, either uniform or non-uniform, by gender or age-group (50 to 64 years versus 65 years and over) (p>0.05) (Table 10.5).

Table 10.5 Differential item functioning in the stair-climbing and walking super-items by gender and age-group in the NorStOP pilot dataset, n=363

Super-item	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Stair-climbing	0.503	1.183 (1)	0.0279	0.685	1.612 (2)	0.2031
Walking	0.327	0.863 (1)	0.3544	0.213	0.561 (2)	0.5717
Age-group						
Stair-climbing	0.084	0.195 (1)	0.6592	0.355	0.820 (2)	0.4424
Walking	0.042	0.111 (1)	0.7393	0.255	0.671 (2)	0.5127

df - degrees of freedom

10.4.8 Targeting

Figure 10.5 shows the locations of persons above the horizontal line (including those with values at the extremes of the scale) and super-item thresholds under the Rasch model below the

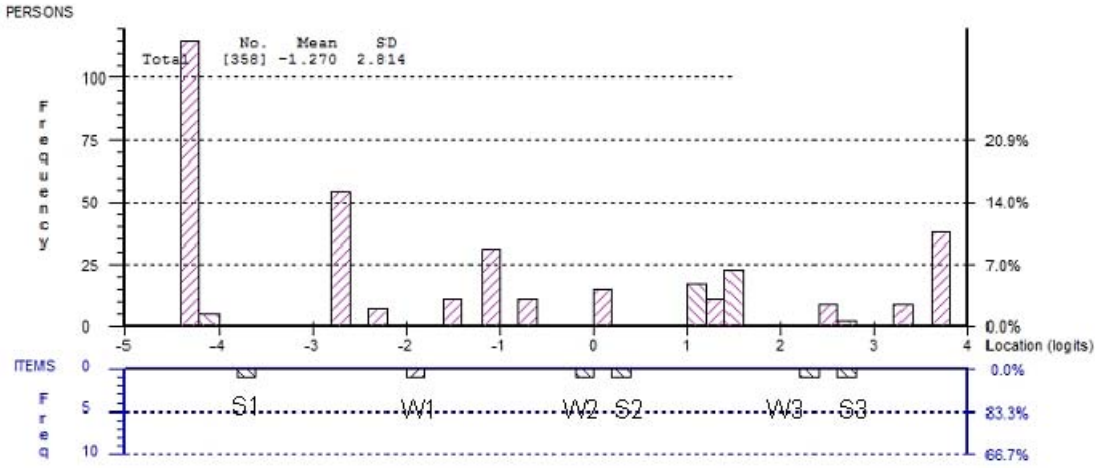
line. A lower score in logits corresponds to a lower level of locomotor disability. There was a spread of super-item thresholds along the continuum of person scores. The first threshold for both of the super-items occurred before the second threshold, which in turn, occurred before the third thresholds.

Approximately 30% of persons have a locomotor disability score below the first threshold, S1, at -3.19 logits. This threshold marks the transition from not being limited in walking more than a mile or climbing several flights of stairs, to not being limited in walking more than a mile and being limited a little in climbing several flights of stairs. The second threshold, W1, is located at -2.28 logits and marks the threshold between not being limited in walking more than a mile and being limited a little in climbing several flights of stairs to being limited a little in both walking more than a mile and climbing several flights of stairs. This process of 'collecting' more disability continues along the x-axis of Figure 10.5. The final item threshold, S3, is located at approximately +2.78 logits. At this point respondents pass from being limited a little in climbing one flight of stairs and limited a lot in walking 100 yards, to being limited a lot in both activities. Approximately 10% of people scored above Threshold S3, and their scores are therefore in the ceiling of the locomotor disability score.

The placement of item thresholds along the locomotor disability continuum gives some idea of the hierarchy of the items. For example, within these items, the first problem that people tend to experience is a little limitation in climbing several flights of stairs; this is followed by a little limitation in walking more than a mile. At the other end of the continuum the last problem within these items that people tend to experience is a lot of limitation in climbing one flight of stairs. This follows very closely after a lot of limitation in walking 100 yards, suggesting that the items are of similar difficulty,

The person separation index of 0.871 provides evidence that the scale scores can be used to distinguish between groups of people (Tennant et al 2008).

Figure 10.5 Person-threshold distribution in the NorStOP pilot dataset, n=363



S1 – Limited a little in climbing several flights of stairs

S2 – Limited a little in climbing one flight of stairs

S3 – Limited a lot in climbing one flight of stairs

W1 – Limited a little in walking half a mile, limited a lot in walking more than a mile

W2 – Limited a little in walking 100 yards, limited a lot in walking half a mile

W3 – Limited a lot in walking 100 yards

10.5 Discussion

10.5.1 Principal findings

In this chapter, an interval-level measure for locomotor disability has been devised from five items from the PF-10 using the Rasch unidimensional measurement model (Rasch 1960). However, the PF-10 items could not be used in their raw state of ordinal responses on three levels, because of response dependency between the items, i.e. the response to one item theoretically implied the response to another and empirical evidence of this hierarchy was provided in Chapter 6. Super-items were created manually using the concept of a Guttman pattern, and in general, these super-items fit the Rasch model reasonably well. There was no differential item functioning, indicating that there was no evidence of item bias due to age or gender.

10.5.2 Strengths and weaknesses of the study

These analyses are unusual for the development of a measure in that they use existing items, i.e. the five items from the PF-10. This is desirable because it does not add to the plethora of health measurement instruments already available (Feinstein et al 1986, de Vet et al 2006).

In general, Rasch analysis can help to avoid the issue of missing data, as scale scores can be computed for those who have completed only some of the items in a scale (Tennant & Conaghan 2007). However, this advantage is limited in the case of the locomotor disability score because of the formation of super-items before the Rasch analysis. Although, a locomotor disability score can be calculated for those people with a score for either the walking or the stair-climbing super-item, there are still some people with missing locomotor disability scores. This is because, in order to get a score for a super-item, a person must complete all of the original PF-10 items that contribute to that super-item. Chapter 6 described the hierarchical nature of the walking and stair-climbing items in the PF-10 and the resulting patterns of missing data, i.e. when a person's response to one item within a set of items implies the responses to others. It was proposed that people do not always complete the other item(s) when they think that their answer implies responses to other items. This is a problem relating to the PF-10 and not directly to the scoring mechanism for this new measure of locomotor disability, but nevertheless deserves consideration.

The use of the Rasch scoring mechanism means that the values assigned to the individual item responses cannot simply be added together, and along with the rescoring of the items to form the super-items, this makes the scoring process more complicated. Although it is difficult to assess the size of the problem this more complicated scoring mechanism may cause, it is likely to be more difficult to use in a clinical than a research setting. It would be possible to use a conversion table between the sum score and interval-level measure, but there are potentially serious problems with this, as discussed in Chapter 9. The main aim of this thesis is to develop a tool for use in epidemiological research, and so the rescoring should not present too great a barrier to its use in this way.

10.5.3 Strengths and weaknesses in relation to other studies

No studies have previously tried to form a measure of locomotor disability from the items of the PF-10. However, other authors have assessed the fit of the 10 original PF-10 items to the Rasch model specification (Haley et al 1994, McHorney et al 1997, Stucki et al 1996, Raczek et al 1998, Jenkinson et al 2001). All of these studies showed there to be some dependency between the items in the PF-10 and concentrating specifically on the five locomotor disability-related items, Haley et al (1994), in their American population, and Raczek et al (1998) in their international populations, found that climbing several flights and one flight of stairs, and walking half a mile and 100 yards showed some dependency. Stucki et al (1996) also found dependency in walking more than a mile in an American population. Jenkinson et al (2001), in their English general population found dependency in only the items relating to climbing one flight of stairs and walking half a mile. Although the authors of these papers claimed that this dependency was not problematic and indeed added to the strength of evidence for unidimensionality, more recent thinking does not support this (Tennant & Conaghan 2007, Marais & Andrich 2008). In this chapter, dependency between the locomotor disability items has been accounted for through the use of super-items. It has also shown more clearly that the items form a single dimension, suitable for forming into a scale. This builds on the work of previous authors who claimed, using Rasch analysis, that the original PF-10 scale was unidimensional, for example Haley et al (1994).

Another instrument used with patients with lower extremity problems is the Lower Extremity Functional Scale (LEFS) (Binkley et al 1999), which is scored via summated ratings, so it does not produce an interval-level measure. However, unlike the locomotor disability measure derived in this chapter, the LEFS was derived from a large bank of items. Through the searching of similar instruments and interviews with patients and clinicians, 75 items were established. Similar items were then merged to create a set of 20 items. The pre-specification of the items in the current locomotor disability tool is a disadvantage and could be seen as a major drawback. However, the tool does cover a range of functional levels and, as shown in this chapter, has good measurement properties.

Jenkinson et al (2001) suggested, in the context of the full PF-10, that computerised adaptive testing (CAT) (Gershon 2004) could be used to reduce the burden placed on patients in

terms of the number of items that they are required to complete. However, this is not possible with the locomotor disability score because of the scoring mechanism with super-items, which is slightly more involved than standard Rasch scoring. However, with only five items to complete, this should not be a cause for concern.

10.5.4 Meaning of the study

The major advantage of having a score created from items that fit the Rasch measurement model is that the score has truly interval-level properties and so mathematical operations can be performed on it in a sensible way (Fischer 1995). This is in contrast to a long ordinal-level scale, such as the PF-10, created by summating ordinal scales, where mathematical operations do not necessarily make sense (see Chapter 2), and it is not possible to test their validity in this context.

10.5.6 Unanswered questions and future research

The creation of this locomotor disability score will allow other studies that have collected data on the PF-10 to assess locomotor disability, hence allowing further research to be carried out without additional respondent burden. The same applies to future studies in which the original PF-10 items can be collected and used to create the locomotor disability score as well as the PF-10 score. Before this measure can be used to further research in the field of locomotor disability, additional testing of the scoring mechanism and its psychometric properites is required. This testing is carried out in the following chapters.

10.6 Conclusions

This chapter has described the development of a scoring mechanism for five items from the PF-10 identified in Chapter 6 as being specific to locomotor disability. The items were used to create super-items that conform to the Rasch model specification and thus provide an interval-level measure. This fit to the Rasch specification ensures that the super-items belong to a single dimension and that interval-level measurement is achieved within the NorStOP pilot study sample. However, it does not mean that the same fit to the Rasch specification would be found in other datasets or that the measure possesses all of the psychometric properties outlined in Chapter 5.

The next chapter assesses the fit of the super-items to the Rasch specification in external datasets, whilst Chapter 12 assesses the psychometric properties of the measure against the desirable characteristics described in Chapter 5.

11 The generalisability the scoring mechanism for the locomotor disability measure

11.1 Introduction

In the previous chapter, an interval-level measure for locomotor disability was derived from the five locomotor disability-specific items of the PF-10 using Rasch analysis. In order to create this measure, the items were required to fit the specification of the Rasch model, as described in Chapter 9. The raw items were not a good fit to the specification, and so super-items were created; one for walking and one for stair-climbing.

The potential problem with this approach is that the success of these super-items in conforming to the Rasch model specification may be specific to the NorStOP pilot study dataset in which they were developed and fitted to the Rasch model. In order to assess the generalisability of the scoring mechanism to create the measure locomotor disability in other samples, it was recreated in several other datasets and the fit of the data to the Rasch model was tested. This chapter describes this testing process, presenting the fit of the super-items, in three datasets external to the NorStOP, to the Rasch model.

11.2 Methods

11.2.1 Datasets

Three datasets were used for the generalisability testing of the locomotor disability measure: The Keele Knee Pain Cohort Study (KNEST) (Jinks et al 2004), The Welsh Health Survey (WHS) (National Assembly for Wales 2000) and a Dutch cohort (van der Windt et al 2008). These three sets of data represent different populations in relation to the NorStOP cohorts. The KNEST study was based in the same English region as the NorStOP cohorts. The data from the WHS was largely representative of the country of Wales, another region of the UK. The Dutch cohort represented a European comparison. All three datasets included the PF-10.

Due to the large size of these three datasets, a random sample of 500 people was selected from each of the cohorts using the statistical analysis programme Stata 9.2 (StataCorp 2005). In general the random samples used for analysis were similar in age and gender structure to the original cohorts from which they were drawn (Table 11.1).

Table 11.1 Cohorts used in the generalisabilty testing of the locomotor disability measure

	KNEST	WHS	Dutch cohort
Location	North Staffordshire	Wales	The Netherlands
PF-10 version used[a]	UK (English)	UK (English 96%) UK (Welsh 4%)	Dutch
Age-group	50 years and over	45 years and over	50 years and over
Sampling frame	3 general practices	Welsh population	5 general practices
Sample size	6,792	17,442	1,112
Female in cohort (%)	55.9	54.3	58.8
Age in cohort, Mean (SD)	65.4 (10.1)	41.7% 65 years and over	64.7 (9.8)
Female in sample (%), n=500	56.2	54.0	61.0
Age in sample, Mean (SD), n=500	65.7 (10.5)	41.6% 65 years and over	64.7 (9.8)

[a]Translations of the PF-10 from English into Welsh and Dutch are given in Appendix G; KNEST - Keele Knee Pain Cohort Study; WHS - Welsh Health Survey; SD - standard deviation

11.2.2 Formation of super-items

The raw PF-10 item scores from each of the samples (KNEST, WHS, Dutch cohort) were arranged into two super-items: stair-climbing and walking, as described in Table 10.2 and entered in to the RUMM2020 package. Categories were then rescored as described in Table 10.3 to create the two super-items, each with four categories.

11.2.3 Testing fit to the Rasch model specification

Criteria for fit to the Rasch model were the same as described in Chapter 9, with one exception. The aim of the analyses carried out in this chapter was to confirm the scoring mechanism used to derive the locomotor disability score. Therefore, Bonferroni corrections were

applied to the p-values for the chi-square and F-tests for individual item fit, the F-tests for DIF and the item-trait interaction chi-square.

Analyses were carried out separately for the three validation datasets: KNEST, WHS and the Dutch cohort.

11.3 Results

11.3.1 The Keele Knee Pain Cohort Study

After the rescoring of categories of the super-items into the four categories obtained in the original derivation of the scoring mechanism, both super-items displayed ordered category thresholds. There was no evidence against the unidimensionality of the scale with 0.08% (95% CI -0.02%, 0.04%) of persons having significantly different locations based on the two super-items. The residual correlation between the two items was -1.00. As in the original testing of the measure in the NorStOP pilot dataset, this was not considered evidence against unidimensionality, as there were only two items in the scale.

Overall item fit was fair, with a mean item residual of 0.551 (SD 0.153). Table 11.2 shows the fit of the individual super-items to the Rasch model specification. The stair-climbing super-item fit well according to all three indices of fit, whilst the walking super-item fit well on the fit residual and the chi-square test after Bonferroni correction. However, there was misfit according to the F-test.

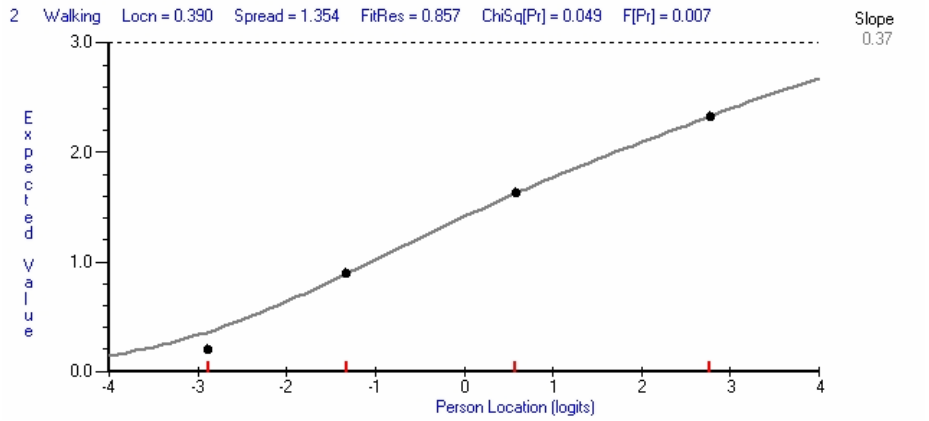
Figure 11.1 shows the item characteristic curve for the walking super-item. It can be seen that the misfit was caused by the lowest class interval and that the data were a good fit to the Rasch model in the other intervals. Hence, this is not a major cause for concern. The mean person fit residual was -0.419 (SD 0.760), showing reasonable fit to the Rasch model. Individual person fit residuals ranged from -1.987 to 0.996. Overall, there was no evidence against invariance of item ordering along the latent trait. The scale was able to measure a higher level of disability than was present in the KNEST sample (mean person location -1.417 (SD 2.697)). However, the PSI was acceptable at 0.845.

Table 11.2 Fit of the super-items in the KNEST sample to the Rasch model

Statistic	Value in the KNEST sample
% of t-tests (95% confidence interval)	0.08% (-0.02%, 0.04%)
Residual correlation (min, max)	-1.00
Mean item fit residual (SD)	0.551 (0.153)
Item residual	
Stair-climbing	0.641
Walking	0.857
Item chi-square test (X^2 (df): p-value)	
Stair-climbing	1.736 (3): 0.6290
Walking	7.859 (3): 0.0490
Item F-test (F (df_1, df_2): p-value)	
Stair-climbing	1.033 (3, 256): 0.3785
Walking	4.116 (3, 242): 0.0072
Mean person fit residual (SD)	-0.419 (0.760)
Person residuals (min, max)	-1.987, +0.996
Item-trait interaction (X^2 (df): p-value)	9.595 (6): p=0.1428
Mean person location (SD)	-1.417 (2.697)
Person separation index	0.845

SD - standard deviation; df - degress of freedom

Figure 11.1 Item characteristic curve for walking super-item in the KNEST sample



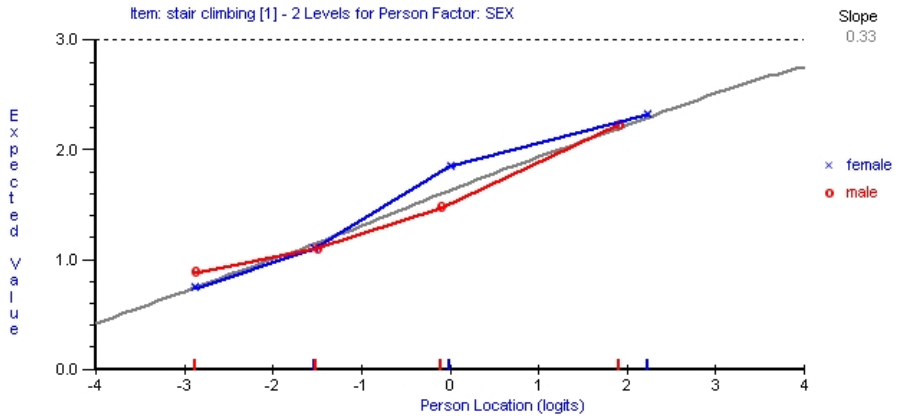
There was no DIF by age-group on either item (p>0.05) (Table 11.3). There was some evidence of non-uniform DIF on the stair-climbing super-item (p=0.0204). This was due to the third class interval (Figure 11.2).

Table 11.3 Differential item functioning in the two super-items by gender and age in the KNEST sample

Super-item	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Stair-climbing	0.175	0.153 (1)	0.6962	1.635	3.323 (3)	0.0204
Walking	0.028	0.057 (1)	0.8120	1.076	2.179 (3)	0.0912
Age-group						
Stair-climbing	0.120	0.236 (1)	0.6277	0.032	0.063 (3)	0.9791
Walking	0.668	1.312 (1)	0.2522	-0.171	-0.338 (3)	0.9999

df - degrees of freedom

Figure 11.2 Differential item functioning by gender in the KNEST sample



11.3.2 The Welsh Health Survey

After the rescoring of categories of the super-items into the four categories obtained in the original derivation of the scoring mechanism, both super-items displayed ordered category thresholds. There was no evidence against the unidimensionality of the scale (Table 11.4). Overall

item fit was fair and both super-items fit well on the fit residual and on the chi-square tests. There was misfit on the F-tests. Figure 11.3 shows the item characteristic curves for the super-items. It can be seen that the misfit was caused by the extreme class intervals. Persons fit reasonably to the Rasch model. There were 24 people with a residual of -2.506. All of these people had not completed one or both of the stair-climbing items on the PF-10 and so their score was based solely on their responses to the walking items. Overall, there was no evidence against the invariance of item ordering along the latent trait, but the scale was able to measure a higher level of disability than was present in the WHS sample. However, the PSI was acceptable at 0.861.

There was no DIF by age-group or gender on either of the super-items (Table 11.5).

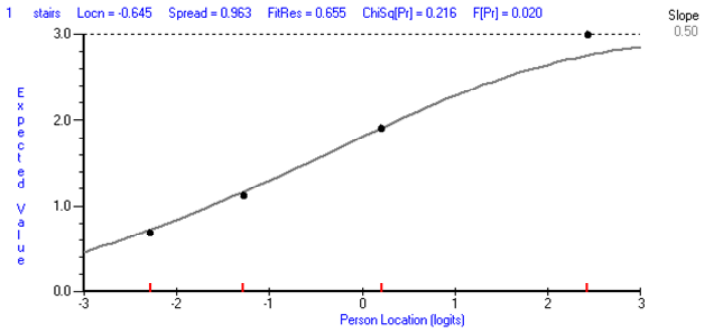
Table 11.4 Fit of the super-items in the WHS sample to the Rasch model

Statistic	Value in the WHS sample
% of t-tests (95% confidence interval)	0.88% (-0.03%, 0.05%)
Residual correlation (min, max)	-1.00
Mean item fit residual (SD)	0.525 (0.183)
Item residual	
Stair-climbing	0.655
Walking	0.396
Item chi-square test (X^2 (df): p-value)	
Stair-climbing	4.459 (3): 0.2160
Walking	6.089 (3): 0.1074
Item F-test (F (df_1, df_2): p-value)	
Stair-climbing	3.412 (3, 121): 0.0197
Walking	4.917 (3, 174): 0.0026
Mean person fit residual (SD)	-0.486 (0.893)
Person residuals (min, max)	-2.506, 0.969
Item-trait interaction (X^2 (df): p-value)	10.548 (6): p=0.1034
Mean person location (SD)	-1.151 (2.548)
Person separation index	0.861

SD - standard deviation; df - degress of freedom

Figure 11.3 Item characteristic curves for stair-climbing and walking super-items in the WHS sample

(a) Stair-climbing



(b) Walking

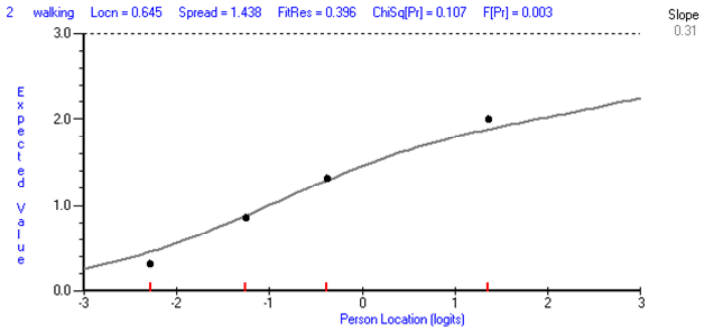


Table 11.5 Differential item functioning in the two super-items by gender and age-group in the WHS sample

Super-item	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Stair-climbing	0.678	1.569 (1)	0.2128	0.163	0.378 (3)	0.7689
Walking	0.553	1.445 (1)	0.2310	0.170	0.445 (3)	0.7212
Age-group						
Stair-climbing	0.001	0.001 (1)	0.9718	1.041	2.504 (3)	0.0626
Walking	1.124	3.037 (1)	0.0827	0.908	2.478 (3)	0.0630

df - degrees of freedom

11.3.3 Dutch cohort

After the rescoring of categories of the super-items into the four categories obtained in the original derivation of the scoring mechanism, both super-items displayed ordered category thresholds. There was no evidence against the unidimensionality of the scale and overall item fit was reasonable. (Table 11.6) shows that there is no misfit of the super-items to the Rasch model. The mean person residual was good, although 12 people had residuals of -2.522. All of these people had not completed one or both of the stair-climbing items on the PF-10 and so their scores were based solely on their responses to the walking items. Overall, there was no evidence against invariance of item ordering along the latent trait. The scale was able to measure a higher level of disability than was present in the WHS sample. However, the PSI remained acceptable.

There was no DIF by age-group or gender in either of the super-items (Table 11.7).

11.4 Discussion

11.4.1 Principal findings

The analyses presented in this section generally supported the fit of the new locomotor disability score to the Rasch model in three external datasets taken from general population samples.

Overall, there was no evidence against unidimensionality and the perfect negative correlation is as might be expected with only two items (David Andrich personal communication, 14 September 2008). The super-items generally fit the Rasch model well with only the F-tests providing evidence against fit. This might also be expected, as F-tests are more sensitive to departures from the Rasch model than the fit residual and the chi-square test. All item misfit appears to be attributable to the class intervals at the extremes of the scales, and when considered on the item characteristic curve, is not representative of large deviations from the model expectations.

Table 11.6 Fit of the super-items in the Dutch sample to the Rasch model

Statistic	Value in Dutch sample
% of t-tests (95% confidence interval)	0.99% (-0.02%, 0.04%)
Residual correlation (min, max)	-1.00
Mean item fit residual (SD)	0.604 (0.869)
Item residual	
Stair-climbing	-0.010
Walking	1.219
Item chi-square test (X^2 (df): p-value)	
Stair-climbing	1.807 (2): 0.4051
Walking	4.831 (2): 0.0893
Item F-test (F (df_1, df_2): p-value)	
Stair-climbing	1.354 (2, 212): 0.2605
Walking	2.150 (2, 227): 0.1189
Mean person fit residual (SD)	-0.374 (0.750)
Person residuals (min, max)	-2.522, 1.184
Item-trait interaction (X^2 (df): p-value)	6.638 (4): p=0.1563
Mean person location (SD)	-1.527 (1.633)
Person separation index	0.764

SD - standard deviation; df - degress of freedom

Table 11.7 Differential item functioning in the two super-items by gender and age-group in the Dutch sample

Super-item	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Stair-climbing	1.115	2.538 (1)	0.1126	0.108	0.247 (2)	0.7817
Walking	0.956	1.708 (1)	0.1926	-0.044	-0.079 (2)	0.9999
Age-group						
Stair-climbing	0.059	0.134 (1)	0.7148	0.8168	1.867 (2)	0.1571
Walking	0.400	0.724 (1)	0.3957	1.0384	1.879 (2)	0.1662

Person fit was generally reasonable. In the WHS and the Dutch datasets, there were several persons with residual fit statistics below the smallest expected value of -2.5 under a Rasch model. In every case, this was because the person had not completed one or both of the original stair-climbing items from the PF-10, and so did not have a score on the stair-climbing super-item. These misfitting persons were not therefore considered to be of concern.

There was no evidence, in any of the three samples, that the assumption of invariance along the latent trait was violated. There was a small amount of evidence of DIF in the KNEST sample, but not in the WHS or the Dutch samples. All three samples had average levels of disability lower than the average level of difficulty of the items, but the score was able to distinguish between groups, as evidenced by the PSI values.

The Dutch cohort was the most different to the NorStOP, and the PF-10 items were taken from a Dutch language translation of the questionnaire. However, the score from this cohort showed the best fit to the Rasch model. This is particularly surprising given that the distances referred to in the Dutch version of the PF-10 are on the metric scale and hence represent different distances: 100 metres = 109.4 yards, half a kilometre = 0.31 miles and a kilometre = 0.62 miles. The KNEST sample, which is the most similar to the NorStOP displayed the worst fit to the Rasch model.

11.4.2 Strengths and weaknesses of the study

The use of pre-existing items to create a new measure of locomotor disability has allowed additional testing of the measure at an early stage. This would not usually have been possible, and is in itself a strength of the study.

Furthermore, the Rasch scoring mechanism has been tested in the three external datasets where the context of the PF-10 in the questionnaire was slightly different each time. This introduced extra variability between the samples in terms of their fit to the Rasch model, and this is to be expected if the locomotor disability score is to be used in different populations.

There are however, ways in which this testing process could have been improved. In particular, the Welsh language version of the PF-10 was completed by 4% of participants in the WHS. This could have resulted in a lack of unidimensionality had people completing the

questionnaire in the two languages responded differently, but this was not the case. It would however have been useful to consider DIF by language in the WHS, but these data were not available. Also, the random sample of 500 chosen from each of the cohort for analysis in this chapter was to an extent arbitrary. This did however, provide for the KNEST and Dutch cohorts around 250 people in which to assess fit to the Rasch specification and approximately 190 people in the WHS having accounted for missing data and extreme scores. This is roughly in line with the sample sizes suggested in the literature (Linacre 1994, Tennant et al 2008).

11.4.3 Strengths and weaknesses in relation to other studies

There have been no previous studies that have aimed to test the scoring mechanism of a new measure developed through the use of Rasch analysis in external samples. This means that there was no precedent for this type of analysis. Previous studies have however, assessed the fit of existing scale items to the Rasch model (for example, Pallant et al (2006), Keenan et al (2007)). Indeed the PF-10 itself has previously been subjected to Rasch analysis (see Chapters 6 and 10). The analyses presented in this chapter were therefore based on these previous studies.

11.4.4 Meaning of the study

The fit of the score to the Rasch model in external datasets is encouraging and provides evidence for the fit of the scoring mechanism to the model, outside the NorStOP cohort. However, this evidence should not be taken as proof that the score is automatically suitable for use in any population. The score is calibrated in a general population, opposed to a patient population, and in those aged 50 years and over, rather than younger people. Hence, despite Rasch parameter estimates being relatively sample-free (Karabatsos 2004), this fit of the data from these general populations to the Rasch model does not necessarily imply that data from more specific, patient populations will also fit the model.

11.4.5 Unanswered questions and future research

As described in the previous chapters, the fit of the super-items to the Rasch model implies that they form a single dimension and result in an interval-level measure. This chapter has shown

that the scoring mechanism derived to form the super-items is generalisable to samples outside the NorStOP pilot dataset. However, there is still the question of whether the measure, derived in the previous chapter, and shown in this chapter to be generalisable, meets the psychometric criteria set out in Chapter 5.

11.5 Conclusion

The measure of locomotor disability derived in the previous chapter has been shown to be generalisable: the method of developing the super-items that fit the Rasch model can be replicated in other datasets. The next chapter tests the new measure against the psychometric criteria set out in Chapter 5, to appraise the suitability of the new measure for the assessment of locomotor disability in large-scale epidemiological studies.

12 Psychometric properties of the locomotor disability measure

12.1 Introduction

The scoring mechanism for the new measure of locomotor disability derived in Chapter 10 has been rigorously assessed in external datasets and shown to be generalisable (Chapter 11). However, the fit of the super-items to the Rasch model, even in a series of datasets, does not constitute full testing of the measure as outlined in Chapter 5. This chapter therefore investigates the psychometric properties of the measure of locomotor disability developed in Chapter 10 in terms of the criteria laid out in Chapter 5.

12.2 Methods

Some of the psychometric properties of the new measure described in Chapter 5 have already been satisfied by the testing of the fit to the Rasch model in Chapter 10 and the assessment of individual items in Chapter 6. Furthermore, the use of the Rasch methodology to create the measure means that some of the techniques to test the measurement properties of the score are different to those described in Chapter 5. Where this is the case, this is stated below.

12.2.1 Datasets

A separate locomotor disability score was deriving using Rasch analysis for each of the cohorts used in this chapter. Where data were longitudinal in nature, i.e. NorStOP pilot study test and retest data (repeatability testing), and NorStOP1 baseline and three-year follow-up data (responsiveness testing), the data were 'stacked', as described in Chapter 9 to ensure that the 'ruler' did not change over time. For simplicity, respondents were only included in longitudinal analyses if they completed at least one super-item at all time points of interest. This need not be the case in other studies, as the Rasch model does not require data at every time point.

12.2.2 Psychometric testing procedures

12.2.2.1 Conceptual and measurement model

The development of this measure of locomotor disability, including the use of the PF-10 items, and in particular the scoring mechanism has been discussed at length in previous chapters. Therefore, this property of the measure will not be discussed further.

12.2.2.2 Unidimensionality

The unidimensionality of the measure was tested and found to be acceptable in Chapter 10. This chapter also reported the PSI for the measure, which is equivalent to Cronbach's alpha. Hence, this property is not considered again in this chapter.

12.2.2.3 Repeatability

The NorStOP pilot study test-retest dataset (see Chapter 4) was used to test the repeatability of the measure over time. For the purpose of these analyses, a stable disability status was defined as choosing the same response option to the SF-12 general health item on the test and re-test questionnaires (interval of 4 weeks apart), as in Chapter 6.

In this section, rather than using the methods outlined in Chapter 5, the stability over time of the locomotor disability measure was assessed using the methods suggested by Hobart & Cano (2009). First, the measure was assessed for equivalence in kind over time, i.e. does the measure work in the same way on both occasions? Second it was assessed for equivalence in degree, i.e. if the measure does work in the same way on both occasions, do people generate equivalent scores? Box 12.1 describes the analyses of equivalence that were carried out.

12.2.2.4 Validity

The face and content validity of the item set have been discussed previously. Hence they are not considered again in this chapter.

Box 12.1 Analysis of equivalence methods

Equivalence in kind

1. A 95% confidence interval was formed around each super-item location estimate at time one (T1). If the locations were equivalent at the two time points, the location of the item at time two (T2) is expected to lie within this interval.
2. The standardised change in the location of each item was calculated

$$\frac{(\text{location}_{T1} - \text{location}_{T2})}{\text{standarderror}(\text{location}_{T1} - \text{location}_{T2})}.$$

Change outside the range -1.96 to +1.96 was considered to show significantly different super-item locations at the two time points.

3. The person locations created from the scale at the two time points were compared using a dependent t-test. A significant result on this test was considered evidence of a change in locations over time.

Equivalence in degree

1. A 95% confidence interval was formed around the person location estimate at T1 for each person. If the locations were equivalent at the two time points, the location of the same person at T2 is expected to lie within this interval.
2. The standardised change in the location of each person was calculated

$$\frac{(\text{location}_{T1} - \text{location}_{T2})}{\text{standarderror}(\text{location}_{T1} - \text{location}_{T2})}.$$

Change outside the range -1.96 to +1.96 was considered to show significantly different person locations at the two time points.

3. A two-way analysis of variance was used to assess DIF by time point, i.e. whether persons responded differently to the items at the two time points despite the same level of locomotor disability. A significant main effect was considered indicative of uniform DIF by time, whilst a significant interaction between time and locomotor disability score was considered indicative of non-uniform DIF by time.

Construct validity

As in Chapter 6, construct validity was assessed by specifying (Box 12.2) and testing hypotheses. The NorStOP1 (Hypotheses 1 to 7) and the CAS-HA (Hypothesis 8) baseline datasets were used in these analyses. Appropriate statistics were used, according to the distribution of the locomotor disability score.

The hypotheses in Box 12.2 were based on those used to test the construct validity of the individual PF-10 items, but were adapted to allow for the locomotor disability score being interval-level rather than a three-level ordinal response. The standards against which to evaluate the locomotor disability measure were the same as previously described in Chapters 4 and 6.

Criterion validity

As discussed in Chapter 6, there is no agreed upon measure of locomotor disability. Therefore, criterion validity was not tested as there was no appropriate gold standard against which to test.

Floor and ceiling effects

Floor and ceiling effects were considered to be present if 15% or more of the sample scored at the extremes of the locomotor disability measure. For the purposes of this measure, in which a higher score represents more disability, in the NorStOP1 baseline sample used for this analysis, a score of -4.26 logits was the ceiling (least disability) and a score of +3.73 logits the floor (most disability). Floor and ceiling effects were calculated in age and gender specific strata.

Interpretation

To a certain extent, the interpretation of the measure has already been considered in the discussion of the association between locomotor disability score and responses to the individual PF-10 items in Chapter 10. In this chapter, normative data are presented from the NorStOP1 subcohort at baseline. To produce the normative scores, the scale score was assessed for Normality and then appropriate statistics used to summarise the scale score overall and by age and gender.

Box 12.2 Hypotheses regarding the construct validity of the locomotor disability measure

1. In those people with self-reported
    a. chest problems;
    b. heart problems;
    c. falls;
    d. breathlessness when walking;
    e. dizziness or unsteadiness;
    f. body mass index of 25 kg/m^2 or greater;
    g. body mass index of less than 18.5 kg/m^2;

    the average locomotor disability score will be higher than in those without these problems.
2. The average locomotor disability score will be higher than in those people
    a. reporting that their general health is “fair” or “poor” than in those reporting their health to be “excellent”, “very good” or “good”.
    b. with any participation restriction, i.e. reporting restriction in one or more of 11 domains, than in those with no restriction.
    c. requiring aids or the assistance of others to move around inside the home than in those who do not require the assistance of others or aids.
    d. requiring aids or the assistance of others go outside than in those who do not require the assistance of others or aids.
    e. who live alone than in those who do not.
3. The average locomotor disability score will be higher than in those who
    a. go out for a walk;
    b. take a bath/shower;
    c. do heavy housework;
    d. do heavy gardening;
    e. do DIY;
    f. walk at least a quarter of a mile;
    g. walk at least two miles;

    less frequently than in those who do these activities more frequently.
4. The average locomotor disability score increase with the number of domains with participation restriction.
5. The average locomotor disability score will increase with higher HADS anxiety scores.
6. The average locomotor disability score will increase with higher HADS depression scores.
7. The average locomotor disability score will increase with higher SIP alertness scores.
8. The average locomotor disability score will increase with lower SPPB scores.

Responsiveness

Responsiveness was assessed using the hypotheses specified in Box 12.3. Differences in mean changes over time between the groups with 95% confidence intervals were used to test these hypotheses. The analyses were carried out in the NorStOP1 cohort using the baseline and three-year follow-up datasets. There is less guidance on the testing of responsiveness than on testing validity. For this reason, the threshold of 75% of hypotheses upheld, as suggested by Terwee et al (2007), was adopted in this chapter to test responsiveness.

The standards against which to evaluate the locomotor disability measure were the same as previously described in Chapters 4 and 6. Onset of a problem or health condition was defined as its absence at baseline and presence at three-year follow-up in those people providing data at both time points. Recovery was defined as the presence of a problem or health condition at baseline and its absence at three-year follow-up.

12.2.2.5 Feasibility

Respondent burden

The respondent burden of the new locomotor disability measure is the same as for completing the original items from the PF-10. This concept will not therefore be considered again in this chapter, except to note that the use of existing items to create the measure will reduce respondent burden relative to creating new items with which to form the measure.

Administrator burden

Administrator burden is discussed from a theoretical perspective.

Missing data

The percentage of missing data for the locomotor disability measure was considered in terms of the number of people without a locomotor disability score in those who responded to the NorStOP1 baseline survey. Percentage of missing data was calculated both overall and in age and gender specific strata.

Box 12.3 Hypotheses regarding the responsiveness of the locomotor disability measure

1. There will be a larger increase in locomotor disability score over three years in those people with
    1. an onset of chest problems, compared to those who remain free of chest problems;
    2. an onset of heart problems, compared to those who remain free of heart problems;
    3. an onset of breathlessness when walking, compared to those who remain free of breathlessness problems;
    4. an onset of dizziness or unsteadiness, compared to those who remain free of dizziness or unsteadiness;
    5. an onset of participation restriction, compared to those who remain free of restriction;
    6. an onset of depression, compared to those who remain free of depression;
    7. an onset of anxiety, compared to those who remain free of anxiety;
    8. begin to live alone, compared to those who continue to live with others;
    9. experience an increase in the number of morbidities they have compared to those who maintain the same number of morbidities;
    10. report worse health than 12 months ago on the SF-36 health transition item compared to those who report staying the same;
    11. report worse health on the SF-12 than they did three years earlier compared to those who report staying the same;
    12. have a change in self-reported weight of 10% or more, compared to those whose weight changes by less than 10%.

2. There will be a larger decrease, or smaller increase, in locomotor disability score over three years in those people who
    6. recover from chest problems, compared to those who continue to have chest problems;
    7. recover from heart problems, compared to those who continue to have heart problems;
    8. recover from breathlessness when walking, compared to those who continue to be breathless;
    9. recover from dizziness or unsteadiness, compared to those who continue to be dizzy or unsteady;
    10. recover from participation restriction, compared to those who continue to have participation restriction;
    11. recover from depression, compared to those who continue to have depression;
    12. recover from anxiety, compared to those who continue to have anxiety;
    13. begin to live with other, compared to those who continue to live alone;
    14. experience an decrease in the number of morbidities they have compared to those who maintain the same number of morbidities;
    15. report better health than 12 months ago on the SF-36 health transition item, compared to those who report staying the same;
    16. report better health on the SF-12 than they did three years earlier, compared to those who report staying the same.

12.3 Results

12.3.1 Repeatability

The locations of the two super-items at Time 1 (T1) and Time 2 (T2) are shown in Table 12.1, along with the standardised change value. The locations for both super-items at T2 are within the 95% confidence intervals of the locations of the super-items at T1. Neither of the standardised change values was outside the range -1.96 to +1.96 again indicating no significant difference in super-item locations between the time periods.

Table 12.1 Equivalence in kind of super- items between the test and retest componenets of the NorStOP pilot study

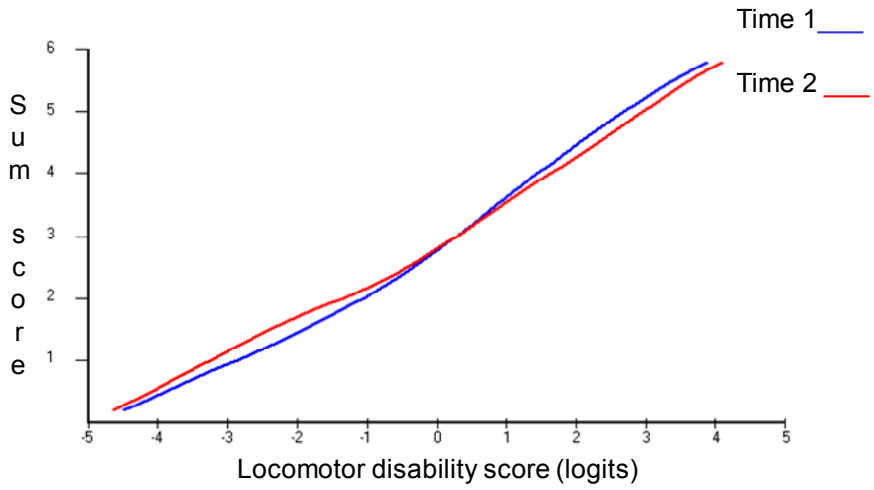
Super-item	Location T1 (95% CI)	Location T2 (95% CI)	Standardised change value (T1-T2)
Stair-climbing	-0.389 (-0.871, 0.093)	-0.366 (-0.86, 0.13)	-0.065
Walking	0.420 (-0.019, 0.859)	0.335 (-0.10, 0.77)	0.268

T1 – Time 1; T2 – Time 2

Figure 12.1 shows the difference in Rasch score locations compared to the sum score between T1 and T2. There is no evidence that the locations are not equivalent in kind ($t=0.385$, $p>0.05$).

Overall, 29 (22.5%) people had locations at T2 outside the 95% confidence interval for their location at T1 and 10 (7.8%) people had a significant difference in their location over the two time points, based on their standardised change scores. The analysis of DIF by time point showed that there was no evidence against equivalence in degree in the individual super-item responses over time ($p>0.05$).

Figure 12.1 Comparison of Rasch score locations and sum scores between the test and retest componenets of the NorStOP pilot study



12.3.2 Validity

12.3.2.1 Construct validity

As the locomotor disability score was not Normally distributed, non-parametric analyses were used to test the construct validity of the score. Medians and interquartile ranges are presented and statistical significance tested using a Kruskall-Wallis test (Vogt 1993). Those people reporting chest problems, heart problems, a fall in the last three months, breathlessness on walking or experiencing dizziness had higher locomotor disability scores than those who did not report these problems (Figure 12.2). Similarly, those who did not report good general health, who had participation restriction in one or more of the 11 KAP areas, and who needed aids or assistance inside or outside the home had higher locomotor disability scores. High (>25kg/m^2) and low (<18.5kg/m^2) BMI and living alone had less pronounced effects, but the difference between groups was still statistically significant ($p<0.001$).

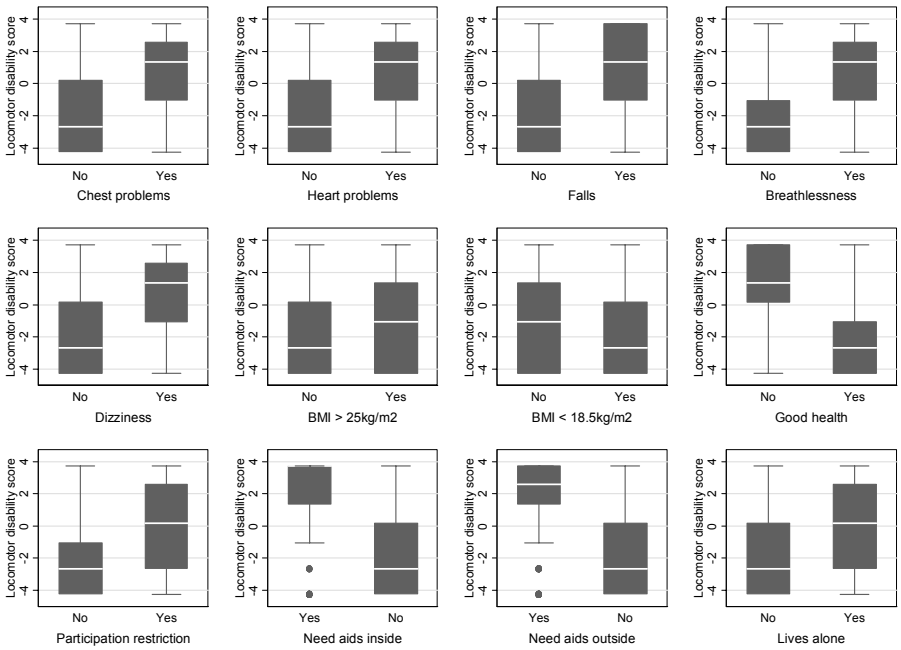
Those doing activities less often had higher locomotor disability scores, whatever activity was considered (Figure 12.3) and these differences were all statistically significant ($p<0.001$).

Those people with more participation restriction had higher locomotor disability scores (Figure 12.4). Similarly, the more likely a person was to be anxious or depressed, the higher their locomotor disability score. Those with a SIP score of zero, indicating a lack of cognitive complaint,

had lower locomotor disability scores than those with higher scores. Those with lower SPPB scores, indicating a higher level of disability in physical performance, had significantly higher locomotor disability scores ($p<0.001$) (Figure 12.5).

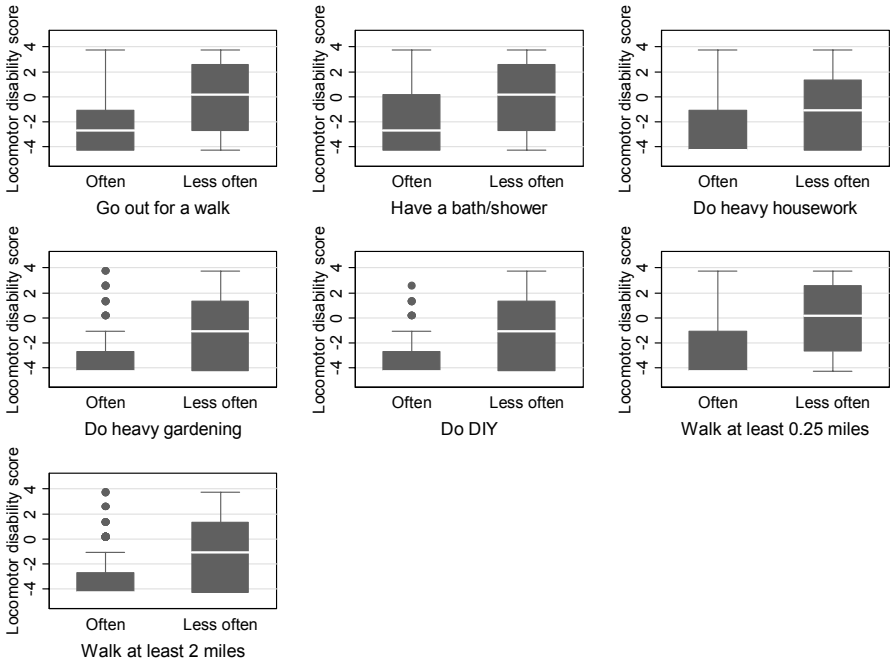
All hypotheses tested in this assessment of construct validity showed statistically significant differences between groups ($p<0.001$) in the direction predicted in Box 12.2. Hence, according to the guidelines of Terwee et al (2007), there is evidence towards the construct validity of the locomotor disability measure.

Figure 12.2 Results of testing the construct validity of the locomotor disability measure: health and socio-demographic variables. NorStOP1



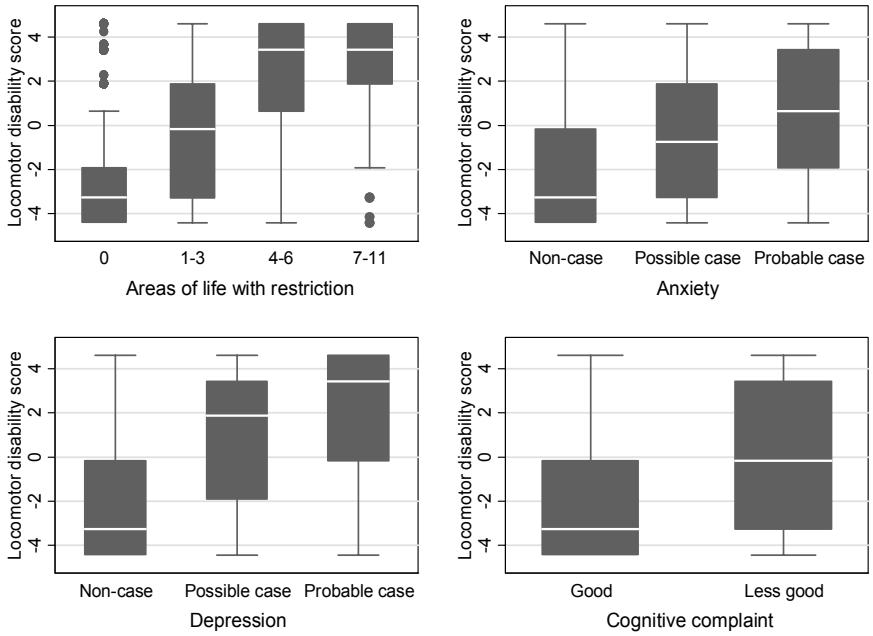
All locomotor disability scores are in logits: a higher score indicates a higher level of locomotor disability. Morbidities and use of aids and assistance self-reported in the NorStOP baseline Health Survey questionnaire (Appendix B, pages 23 and 10 respectively). General health assessed by the SF-12 (Ware et al 1996); participation restriction assessed by the KAP (Wilkie et al 2005).

Figure 12.3 Results of testing the construct validity of the locomotor disability measure: frequency of activities. NorStOP1



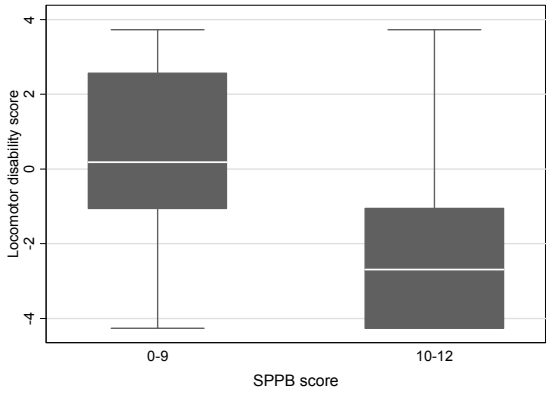
All locomotor disability scores are in logits: a higher score indicates a higher level of locomotor disability. All activities self-reported in the NorStOP baseline Health Survey questionnaire (Appendix B, pages 8 and 9).

Figure 12.4 Results of testing the construct validity of the locomotor disability measure: comparison to other scales. NorStOP1



All locomotor disability scores are in logits: a higher score indicates a higher level of locomotor disability. Participation restriction assessed by the KAP (Wilkie et al 2005); Anxiety and depression assessed by the HADS (Zigmond & Snaith 1983); Cognitive complaint was assessed by the SIP Alertness subscale (Bergner et al 1981).

Figure 12.5 Results of testing the construct validity of the locomotor disability measure: Short Physical Performance Battery. CAS-HA



All locomotor disability scores are in logits: a higher score indicates a higher level of locomotor disability. SPPB - Short Physical Performance Battery (Guralnik et al 1994)

12.3.2.2 Floor and ceiling effects

Table 12.2 shows the floor and ceiling effects present in different age and gender groups. There was a substantial ceiling effect in all except the oldest age-groups and the effects were more pronounced in males than in females. There was a floor effect in males aged 80 years and over and in females age 70 years and over.

Table 12.2 Floor and ceiling effects in the locomotor disability score by age-group and gender (%)

%	Males		Females		Overall	
	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling
50 to 59 years	5.0	56.5	5.4	42.1	5.2	48.9
60 to 69 years	8.6	40.3	8.1	29.4	8.4	34.5
70 to 79 years	11.0	24.8	15.0	15.7	13.2	19.7
80 years and over	22.7	12.6	31.9	4.8	28.6	7.6
Overall	9.2	39.7	12.1	269	10.8	32.7

Floor: highest possible score (-4.26 logits), i.e. most disability on the measure; Ceiling lowest possible score (3.73 logits), i.e. least disability on the measure

12.3.2.3 Interpretation

The locomotor disability score was not Normally distributed in the NorStOP1 at baseline. Table 12.3 shows the median and inter-quartile ranges of the scale scores by age-group and gender. Scores were higher, indicating more locomotor disability, in females and those who were older at baseline.

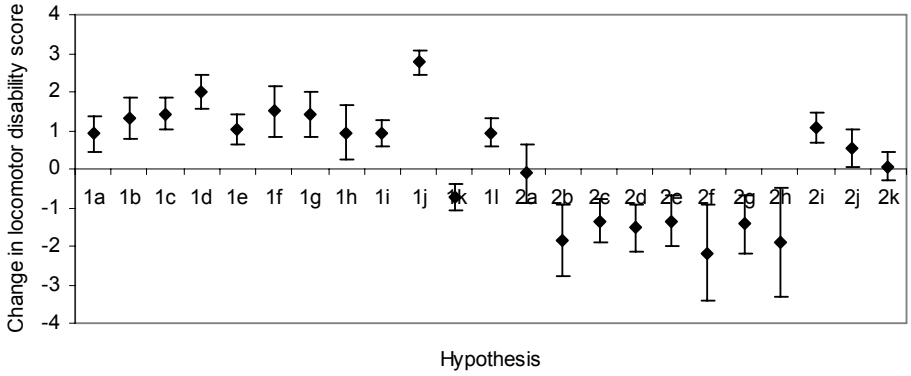
Table 12.3 Median (IQR) locomotor disability scores (logits) by age and gender

Median (IQR)	Males	Females	Overall
50 to 59 years	-4.26 (-4.26, -1.06)	-2.69 (-4.26, 0.18)	-2.69 (-4.26, -1.06)
60 to 69 years	-2.69 (-4.26, 0.18)	-1.06 (-4.26, 0.18)	-2.69 (-4.26, 0.18)
70 to 79 years	-1.06 (-2.69, 1.35)	0.18 (-2.69, 2.57)	0.18 (-2.69, 1.35)
80 years and over	1.35 (-1.06, 2.57)	2.57 (0.18, 3.73)	1.35 (0.18, 3.73)
Overall	-2.69 (-4.26, 0.18)	-1.06 (-4.26, 1.35)	-1.06 (-4.26, 1.35)

12.3.2.4 Responsiveness

Twenty-three hypotheses were tested to assess the responsiveness of the locomotor disability measure. Eighteen (78.3%) of these were upheld (Figure 12.6). This satisfies the criterion of 75% suggested by Terwee et al (2007) to be necessary for validity, and which has been adopted in this study for responsiveness.

Figure 12.6 Results of responsiveness analysis: unadjusted mean change in locomotor disability score over three years (95% confidence interval). NorStOP1



1a - onset of chest problems; 1b - onset of heart problems; 1c - onset of breathlessness; 1d - onset of dizziness or unsteadiness; 1e - onset of participation restriction; 1f - onset of depression; 1g - onset of anxiety; 1h - begin to live alone; 1i - increase in the number of morbidities; 1j - worse health than 12 months ago; 1k - worse health than three years earlier; 1l - weight change of at least 10%; 2a - recover from chest problems; 2b - recover from heart problems; 2c - recover from breathlessness; 2d - recover from dizziness or unsteadiness; 2e - recover from participation restriction; 2f - recover from depression; 2g - recover from anxiety; 2h - begin to live with other; 2i - decrease in the number of morbidities; 2j - better health than 12 months ago; 2k - better health than three years earlier

12.3.3 Feasibility

12.3.3.1 Administrator burden

The burden on the administrator in terms of data collection is the same for the new measure of locomotor disability as for the original items from the PF-10. However the creation of the scale score is more complex and requires specialist knowledge and computing software.

It is possible to create the locomotor disability score by creating a sum score and a conversion table to get a score in logits. However, as discussed in Chapter 10, this is not recommended and may result in a score that is not on an interval-level. In order to create the

measure from the raw PF-10 items, as implemented in this chapter, access to a Rasch analysis computer package, such as RUMM2020, is required. This can be costly and requires knowledge of how to use the package properly.

Despite these burdens on the administrator, the new measure of locomotor disability is much more flexible in terms of analysis than the scores derived from summated ratings and other similar methods that do not result in interval-level measurement: allowing a greater range of analyses to be undertaken. The use of Rasch analysis also permits the anchoring of the measure over time to guard against the 'ruler' changing over time (Wright 2003).

12.3.3.2 Missing data

Table 12.4 shows the level of missing data overall and in age and gender specific strata. Levels of missing data were higher in females and at older ages. Overall, approximately 12% of people in the NorStOP1 baseline sample do not have a locomotor disability score.

Table 12.4 Percentage of individuals without a locomotor disability score by age and gender

%	Males	Females	Overall
50 to 59 years	7.1	8.7	7.9
60 to 69 years	8.0	11.9	10.1
70 to 79 years	12.6	17.4	15.4
80 years and over	16.5	22.4	20.4
Overall	9.7	13.9	12.1

12.4 Discussion

12.4.1 Principal findings

The measurement properties of the score were tested in the NorStOP pilot, the NorStOP1 baseline and three-year follow-up datasets and the CAS-HA dataset. The locomotor disability measure was shown to be reproducible over a four-week period with the ability to distinguish

between people, despite any measurement error. There was no differential item functioning over time. The score was shown to be highly reproducible in kind, i.e. the scale worked in the same way over time, and acceptably reproducible in degree, i.e. given the scale worked in the same way over time, people generated equivalent scores, over a four-week period.

Construct validity of the measure was examined by testing a series of hypotheses and good evidence of this validity was found.

The score was shown to have substantial ceiling effects in younger age-groups and floor effects in the older age-groups. Overall, a score could not be calculated for 12% of people, but this rose to approximately 30% in those aged over 80 years, as described for the individual items in Chapter 6.

The responsiveness of the score to changes in the construct of locomotor disability was also examined using hypotheses, of which 78% were upheld. This provides evidence in support of the responsiveness of the measure according to the criteria of Terwee et al (2007) for validity, although no such criteria exist for responsiveness and the tests were based on statistical significance, rather than specific effect sizes. The presence of a large ceiling effect might have been expected to limit responsiveness, but there was little evidence of this.

12.4.2 Strengths and weaknesses of the study

These analyses are unusual for the development of a measure in that they use existing items, i.e. the five items from the PF-10. This allowed the evaluation of the measurement properties of the score in the NorStOP1 cohort. However, restricting the items in the scale to those from the PF-10 has to some extent limited the content validity of the scale and is likely to be the cause of the ceiling and floor effects in the score. This is should not be practically problematic if one is willing to assume that those who do not cross the first threshold on the scale, i.e. can climb several flights of stairs and walk more than a mile with no limitation, are mobile enough that this does not affect their day to day lives and that those who cross the last threshold, i.e. are limited a lot in climbing one flight of stairs and walking 100 yards, have substantial, clinically important disability. Aside of these practical issues though, there are methodological problems with the presence of such effects. Cross-sectionally, people within the ceiling or the floor cannot be distinguished from one another

because they are all given the same logit score. Longitudinally, people in the floor or the ceiling can only change in one direction on the scale, i.e. those at the floor can only improve (decrease their score) and those at the ceiling can only deteriorate (increase their score), so deterioration in those at the floor or improvement in those at the ceiling will not be detected: this can result in the phenomenon of regression to the mean. Furthermore, changes in score involving the ceiling or floor are difficult to interpret because all people in the group receive the same logit score, estimated from the geometric mean of the three easiest or most difficult items completed (Andrich et al 2003), which may not be representative of their actual level of disability.

In general, Rasch analysis can help to avoid the issue of missing data, as scale scores can be computed for those who have completed only some of the items in a scale (Tennant & Conaghan 2007). This advantage is limited in the case of the locomotor disability score because of the formation of super-items before the Rasch analysis, and resulted in a large proportion of those in the oldest age-groups not having a locomotor disability score. This is primarily a criticism of the original PF-10, but it has consequences for the use of this new measure of locomotor disability.

The evidence in favour of the responsiveness of the measure is fairly limited in nature, being adapted from analysis of construct validity. This is a weak point of the psychometric testing procedures employed in this thesis. Hypotheses may have been incorrectly specified. This seems unlikely, given the similarity with the hypotheses for construct validity that were also upheld. There is however some evidence, that despite its importance cross-sectionally (Iezzoni et al 2001), living arrangement does not affect changes in functioning longitudinally (Michael et al 2001).

Furthermore, the presence of floor and ceiling effects and the short length of the locomotor disability scale make the evidence for responsiveness surprising. With only seven points on the measure a fairly large change in locomotor disability is needed to move between points on the measure. Therefore, the score would be thought to be less sensitive to change than a score with more points. Furthermore, the data used for the responsiveness analyses in this study were from two surveys three years apart, which could be considered to be too long to draw firm conclusions regarding responsiveness. It would likely be worthwhile employing more sophisticated analysis techniques to these data as and when they become available.

12.4.3 Strengths and weaknesses of the study in relation to the literature

In combination with earlier parts of this thesis, this chapter has provided a comprehensive review of the psychometric properties of the new measure of locomotor disability developed in Chapter 10. With few exceptions, the psychometric properties described in Chapter 5 have been addressed. Exceptions to this were criterion validity and minimal important change, as no suitable benchmarks against which to assess the measure were available in the NorStOP. This is therefore one of the most comprehensive evaluations of any instrument intended for the assessment of locomotor disability, as the majority of previous studies have used unrecognised items for their assessments (see Chapter 3). Furthermore, the testing of this new measure has used only methods of analysis suitable for the level of measurement of the score and its distribution.

12.4.4 Meaning of the study

The new measure of locomotor disability as been shown to have good measurement properties. It is therefore a suitable candidate for inclusion and use in future studies. It should however be noted that the measure was developed in general population data from a single region of the UK, and although the scoring mechanism was shown to be generalisable to other datasets (Chapter 11), it should be fully tested before it is used in other settings.

12.4.5 Unanswered questions and future research

In order to be truly useful in longitudinal studies, a minimal important change (MIC) is needed to identify how big a change in score is a meaningful change (Terwee et al 2007). This MIC will now make sense, because of the interval-level nature of the score, where a change of, say 0.8 logits, is equivalent at different points along the locomotor disability continuum. Current thinking suggests that the best way to create an MIC is through a combination of two traditional methods: an anchor-based approach and a distribution-based approach (Crosby et al 2003, de Vet et al 2007). In order to use this method, a suitable anchor is required against which to compare the locomotor disability measure. Unfortunately, such a standard was not available in the NorStOP. Furthermore, there is no obvious standard that should be used in considering meaningful change in

locomotor disability; a combination of self-reported and observed assessments might be useful to give a full picture.

Further work might also include the investigation of substituting missing values in the original PF-10 items where the score is obvious from the Guttman pattern of the items. This might reduce the level of missing data in studies where item response rates are problematically low.

12.5 Conclusions

It is possible to create a measure of locomotor disability from pre-existing items of the PF-10 that is generalisable to other population samples and has good psychometric properties. It remains to demonstrate that this measure is necessary to the future of research in the field of locomotor disability and that it can provide additional information over and above that provided by a dichotomous definition. The next chapter describes methods of analysing interval-level data in a longitudinal setting, before data from the NorStOP1 over the full six years of follow-up are used to demonstrate some of the potential benefits of this interval-level measure.

13 Approaches to the analysis of an interval-level measure over time

13.1 Introduction

Chapters 9 to 12 described the development and testing of a new scoring mechanism for items from the PF-10 to create an interval-level measure of locomotor disability. This measure has the potential to provide more information about locomotor disability at the population level, particularly in longitudinal studies. For example, rather than simply estimating changes in prevalence, and the rate of the onset of and recovery from locomotor disability, it is possible to look at the change in the whole distribution of locomotor disability in the population of interest in the manner suggested by Rose (1985), as well as summarising this change in terms of the average rate of change within that population.

Further to this, an interval-level measure allows the investigation of factors associated with higher rates of change in an outcome, for example locomotor disability, rather than simply factors associated with the onset of or recovery from that outcome. It is possible that these factors are different and that consideration of only binary definitions of the outcome will miss important associations.

This chapter illustrates some potential methods for analysing an interval-level measure over time.

13.2 Potential methods for the longitudinal analysis of interval-level measures

This section will outline some potential methods available for the analysis of interval-level measures longitudinally in large-scale population epidemiology studies. Two major methods will be

considered: graphical techniques, to compare distributions over time; and regression models, to assess associations between changes in the outcome and independent factors.

13.2.1 Changes in group level characteristics over time

When considering interval-level data and potential methods to analyse those data, it is imperative to fully understand the data. This includes the distribution of the data and the way in which they behave. Graphical methods of data analysis can be enlightening for these purposes, and this section describes some of the methods that can be used to graphically investigate interval-level data.

13.2.1.1 Kernel density function plots

A kernel density function is a non-parametric method of estimating a probability density function from a set of observed data. This function can be displayed graphically to visualise the distribution of the data as a smooth curve. Curves can be overlaid to compare distributions. If two distributions are identical, then the kernel density functions will lie exactly on top of one another. If the population providing the second distribution has experienced a right (upward) shift compared to the first population, then the whole kernel density curve will move to the right (Flegal & Troiano 2000) (Figure 13.1a). Similarly if the second distribution experiences a left (downward) shift, the curve will move to the left. If the second distribution is more skewed than the first, this can also be seen by overlaying the kernel density functions (Figure 13.1b).

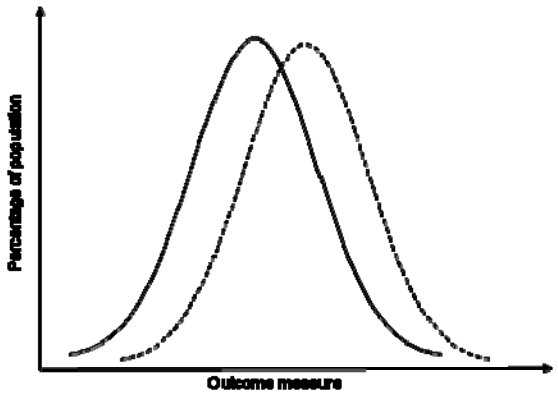
The process of fitting a kernel density function can be envisaged as a series of histograms. In a standard histogram, the continuum of the measure is divided into intervals (or bins) and the frequency of observations within each interval is plotted at the centre of the interval in a bar chart. A kernel density estimate is created from what can be envisaged as a series of histograms with overlapping intervals. Furthermore, rather than simply counting the frequency of observations in an interval the kernel density function assigns weights to the observations, dependent on their proximity to the centre of the interval. These weighted values are then summed. There are several methods by which to create a kernel density function: methods vary according to the type of kernel used to determine the weights for the observations. A kernel is a smooth, symmetrical probability

density function that integrates to one (Salgado-Ugarte et al 1993). There are several kernel functions available; each has a different efficiency, which is a trade-off between variance and bias.

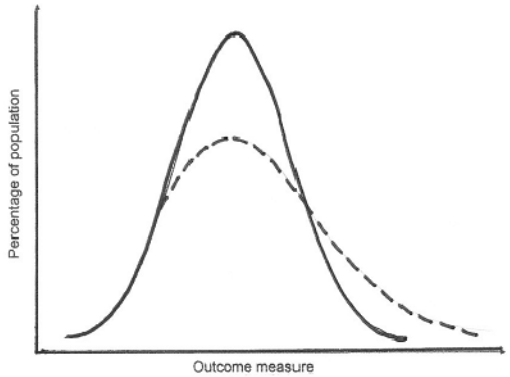
The Epanechnikov kernel function (Epanechnikov 1969) is maximally efficient (Salgado-Ugarte et al 1993). The process of fitting and plotting a kernel density estimate results in a smooth curve that is not dependent on the choice of intervals, as a histogram would be. Kernel density plots also have the advantage than more than two comparisons can be made, as more than two plots can be overlaid on the same graph.

Figure 13.1 Schematic representation of kernal density plots

a. Constant right shift in distribution



b. Increase in skewness in the distribution



13.2.1.2 Cumulative density function plots

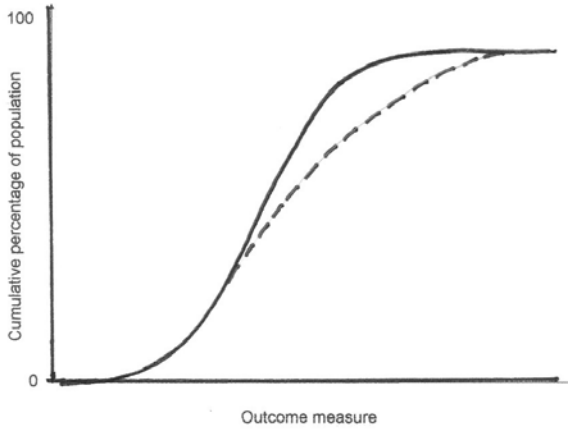
A cumulative density function describes the probability that an observation will be less than a given value of an interval-level measure. For example, in the locomotor disability measure derived in this thesis, one might wish to calculate the probability of a score of less than +2 logits. Given the cumulative density function, one could calculate this probability.

Cumulative density functions can be plotted and used to compare the distribution of a particular variable across populations. As with the kernel density plots, a constant right (or left) shift in the second distribution compared to the first results in a parallel curve (Figure 13.2a). An increase in skewness in the second distribution compared to the first is seen in a cumulative density plot in the form of a deviation from the density of the first distribution at the higher end of the distribution (right/upward skew) (Figure 13.2b). Like the kernel density plots, cumulative density plots have the advantage that more than two distributions can be compared simultaneously.

Figure 13.2 Schematic representation of cumulative density plots

a. Constant right shift in distribution

b. Increase in skewness in the distribution



13.2.1.3 Tukey mean-difference plots

Tukey mean-difference plots compare the quantiles of two distributions from a measure. They plot the mean value of the measure in each quantile on the x-axis against the difference in the mean values within that same quantile on the y-axis. If the two distributions are identical, the plot will be a straight horizontal line at a value of zero on the y-axis, any deviation from this pattern shows that the distributions are different. A movement of the straight horizontal line up from zero indicates a right (upward) shift in the second distribution compared to the first (Figure 13.3a), whilst a movement down from zero indicates a left (downward) shift. A deviation from a straight horizontal line indicates an increase in skewness in the second distribution compared to the first (Figure 13.3b). It is also possible to have both of these situations occur at the same time. This would result in a curved line that was shifted above or below a y-value of zero. Unlike the kernel and cumulative density plots, Tukey mean-difference plots only allow the comparison of two distributions at one time.

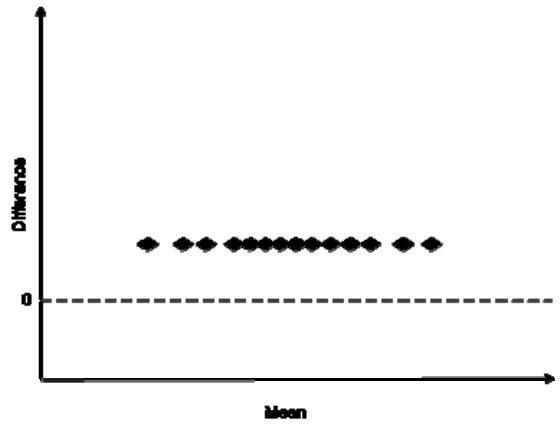
13.2.2 Associations between changes in outcome over time and independent factors

Whilst some research questions call for the simple consideration the changes in the distribution of an outcome over time, in some circumstances a question can only be answered by quantifying the association between changes in an individual and the factors to which that individual has been exposed. This section considers methods suitable for quantifying changes over time on an interval-level outcome and their association with other factors.

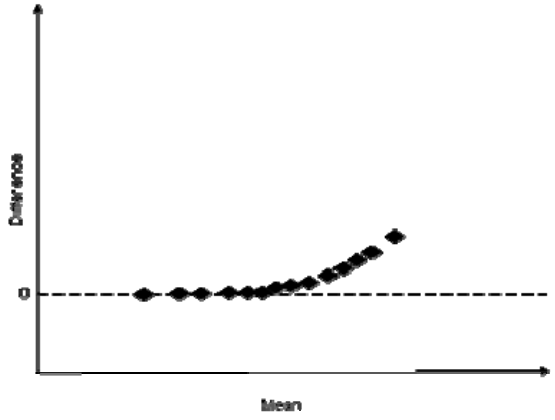
In quantifying the change over time in an individual it is necessary to take into account the correlation between observations: measurements on the same person are likely to be more highly associated with each other than measurement on different individuals. The methods considered below are all able to account for this correlation.

Figure 13.2 Schematic representation of Tukey mean-difference plots

a. Constant right shift in distribution



b. Increase in skewness in the distribution



13.2.2.1 Paired t-tests

A simple method to quantify the difference between measurements at two time points is the use of a paired t-test. The paired nature of the test accounts for the correlation between observations within an individual over time, and it can be employed when the distribution of the differences between the time points is approximately Normal, or when the sample size is greater than 25 (Twisk 2003). The advantage of this approach is that it allows a confidence interval, and hence some easily understood estimate of uncertainty around the estimate to be calculated. However, the method can only be used to assess the course of the outcome between two time points.

13.2.2.2 Multivariate analysis of variance

Multivariate analysis of variance (MANOVA) can be viewed as an extension to the paired t-test, in that it can quantify changes in the course of an outcome over more than two time points (Twisk 2003). This approach also allows comparisons of this course to be made between two or more groups within the sample, for example males and females, or those reporting pain in the knee, hip or foot, say.

Despite these advantages of the MANOVA approach in comparison to the paired-t-test, it does have some restrictions. First, the outcome of interest must be multivariately Normally distributed (although in large samples, MANOVA is fairly robust to this assumption and an non-parametric equivalent test is available in the form of the less powerful Friedman test). Second, only those individuals with outcome data at all time points can be included in the analysis and MANOVA cannot take account of different time intervals between measurements. Third, only p-values result from this method: it does not provide an easily interpretable quantification of the course of change in the outcome. Finally, although MANOVA can assess the differences in the course of an outcome between groups, which the paired t-test cannot, it can only consider one grouping variable and this variable must be categorical in nature.

13.2.2.3 Change score methods

A widely used group of methods for assessing interval-level measures in a longitudinal setting is the calculation and analysis of change scores (Twisk 2003). Like the paired t-test described above, these methods can only be used to assess the change in the outcome between two time points. Nevertheless they are popular in the literature and have received a great deal of attention.

A change score is usually calculated by subtracting the score on the outcome at the second time point from the score at the first time point. It has also been suggested that the score could be transformed to be a relative difference by dividing by the score at the first time point to obtain a relative change (Twisk 2003). This simple calculation of a change score has been criticised when the measure used as the outcome is bounded, as is the case with most scores from health assessment instruments. Such measures are liable to the presence of ceiling and floor

effects, which can lead to an ill-defined change score, as it is not known by how much individuals in the floor and ceiling have changed. It has therefore been suggested that a correction be made to the definition of change in the presence of these floors or ceilings, for example,

when $Y_{it2} > Y_{it1}$: $\Delta Y = \frac{Y_{it2} - Y_{it1}}{Y_{max} - Y_{it1}}$;

when $Y_{it2} < Y_{it1}$: $\Delta Y = \frac{Y_{it2} - Y_{it1}}{Y_{it1} - Y_{min}}$;

when $Y_{it2} = Y_{it1}$: $\Delta Y = 0$;

where Y_{it1} are the observations for subject *i* at time t_1, Y_{it2} are the observations for subject *i* at time t_2, Y_{max} is the maximum possible value of *Y,* and Y_{min} in the minimum possible value of *Y* (Twisk 2003).

However, it is unclear what course of action one should take in the presence of both a floor and a ceiling effect, where such formulae fail to correct for the problem (Twisk 2003). The major limitation of the use of a change score is the statistical phenomenon of regression to the mean, whereby individuals with high scores at the first time point will by chance be expected to have lower scores at the second time point, whilst individuals with a low score at the first time point will be expected to have a higher score at the second time point (Barnett et al 2005).

Changes scores can be modelled using ordinary least squares regression. A change score model is therefore easily fitted in all commercially available software packages, and the results of the model easily understood and interpreted.

Much has been written about adjustment for baseline values in the analysis of change scores with the objective of eliminating confounding, or removing bias due to ceiling or floor effects (Glymour et al 2005). In such a model, the score at the first time point (y_1) has already been included in the calculation of the change score, i.e. y_1-y_2. Glymour et al (2005) showed using directed acyclic graphs that the 'double counting' of the score at the first time point, when adjustment is made for baseline score, can lead to inflated estimates of the regression coefficients because the errors for the first time point and change scores are correlated. It is not therefore recommended to make adjustments for baseline scores in a change score model.

An alternative to the explicit calculation of a change score is to use the outcome measure at the second time point as the dependent variable, with value of the measure at the first time point as an independent variable in the model. This approach is often known as 'analysis of covariance' and

can ‘correct’ for the phenomenon of regression to the mean seen in the analysis change scores (Twisk 2003). However, in the presence of a floor or ceiling effect, the assumption of Normally distributed errors may not be met.

Whether a change score is calculated directly, or an analysis of covariance is use to model the same effect, there is the advantage over the methods discussed above that adjustment can be made in the regression model for the effects of other variables. Furthermore, differences in the course of change can be considered in terms of continuous rather than only categorical variables of interest.

13.2.2.4 Generalised estimating equations

Generalised estimating equations (GEE) model the association between a set of independent variables and the outcome measure at all time points simultaneously. Using an iterative quasi-likelihood procedure, regression coefficients are estimated that represent the longitudinal development of the outcome in relation to the independent variables. These coefficients can be difficult to interpret, as they combine the effect of an independent variable, say gender over time within an individual and between individuals at a single time point.

It is also necessary when using a GEE approach to specify the level of correlation in the outcome that is expected between time points. This is known as the ‘working correlation matrix’. There are several types of working correlation structure, as outlined in Box 13.1. There is no simple way to chose which of these correlation structures to use, and this choice can influence the conclusions of analysis (Twisk 2003). One way that has been suggested to choose the appropriate matrix is to follow a three-step procedure (Twisk 2003): i) estimate a naive linear regression model (assuming independence over time points); ii) use the residuals of this model to estimate a correlation structure; iii) re-estimate the regression coefficients, specifying the correlation matrix as suggested by ii).

Although it may be difficult to decide upon the appropriate correlation structure to adopt, and regression coefficients have a mixed interpretation, the GEE approach offers two advantages that may be useful together or alone: it is possible to allow for unevenly spaced time intervals and independent variables can be allowed to vary over time.

Box 13.1 Generalised estimating equations: work correlation structures

Independent structure

Correlations between measurements of the outcome over time are assumed to be zero. For example, with five time points:

	t1	t2	t3	t4	t5
t1	-	0	0	0	0
t2	0	-	0	0	0
t3	0	0	-	0	0
t4	0	0	0	-	0
t5	0	0	0	0	-

Exchangeable structure

Correlations between measurements of the outcome over time are assumed to be equal, irrespective of the length of time between measurements. For example, with five time points:

	t1	t2	t3	t4	t5
t1	-	ρ	ρ	ρ	ρ
t2	ρ	-	ρ	ρ	ρ
t3	ρ	ρ	-	ρ	ρ
t4	ρ	ρ	ρ	-	ρ
t5	ρ	ρ	ρ	ρ	-

m-dependent structure

Also called the stationary structure. Correlations one measurement apart are equal at a value of ρ_1. Correlations two measurements apart are equal at a value of ρ_2, and so on, to correlations m measurements apart at a value of ρ_m. Correlations more than m measurements apart are assumed to be zero. For example, with five time points and m=2:

	t1	t2	t3	t4	t5
t1	-	ρ_1	ρ_2	0	0
t2	ρ_1	-	ρ_1	ρ_2	0
t3	ρ_2	ρ_1	-	ρ_1	ρ_2
t4	0	ρ_2	ρ_1	-	ρ_1
t5	0	0	ρ_2	ρ_1	-

Autoregressive structure

Correlations one measurements apart are assumed to be ρ. Correlations t measurements apart are assumed to be ρ^t, for t=2, 3, ... For example, with five time points:

	t1	t2	t3	t4	t5
t1	-	ρ	ρ^2	ρ^3	ρ^4
t2	ρ	-	ρ	ρ^2	ρ^3
t3	ρ^2	ρ	-	ρ	ρ^2
t4	ρ^3	ρ^2	ρ	-	ρ^1
t5	ρ^4	ρ^3	ρ^2	ρ	-

Unstructured

All correlations are assumed to be different. For example, with five time points:

	t1	t2	t3	t4	t5
t1	-	ρ_1	ρ_2	ρ_3	ρ_4
t2	ρ_1	-	ρ_5	ρ_6	ρ_7
t3	ρ_2	ρ_5	-	ρ_8	ρ_9
t4	ρ_3	ρ_6	ρ_8	-	ρ_{10}
t5	ρ_4	ρ_7	ρ_9	ρ_{10}	-

13.2.2.5 Random coefficients models

The random coefficients model is also known as a random effects, multilevel or mixed effects model. As with the GEE approach, these models can describe the association between a set of independent variables and the outcome measure at all time points simultaneously.

Random coefficient models correct for the correlation of observations within groups. In the case of a longitudinal dataset, these groups are individuals: observations over time are grouped within an individual. It is also possible within this framework to allow for the clustering of individuals, for example patients within a general practice.

The principle of a random coefficient model is that the regression coefficients for each individual are modelled separately. Hence, in the simplest form of random effects model, the random intercept model, the intercept parameter varies between individuals. Hence the regression equation is, $Y_{it} = \beta_{0i} + \beta_1 t + \varepsilon_{it}$, where Y_{it} are the observations of individual i at time t, β_{0i} is the random intercept for individual i, β_1 is the regression coefficient for time and ε_{it} is the error for

individual *i* at time *t*. The model is equivalent to a GEE approach using an exchangeable correlation structure (Twisk 2003). The next simplest model, the random slope model, allows the slope to vary between individuals. The equation for this model is, $Y_{it} = \beta_0 + \beta_{1i}t + \varepsilon_{it}$, where Y_{it} are the observations of individual *i* at time *t*, β_0 is the intercept, β_{1i} is the random regression coefficient for time and ε_{it} is the error for individual *i* at time *t*.

A more complex model can have both the intercept and the slope vary for each individual. The regression equation then becomes, $Y_{it} = \beta_{0i} + \beta_{1i}t + \varepsilon_{it}$, where Y_{it} are the observations of individual *i* at time *t*, β_{0i} is the random intercept, β_{1i} is the random regression coefficient for time and ε_{it} is the error for individual *i* at time *t* The results of the modelling process are presented as the mean of the coefficients for individuals.

The main assumption of this model is that the so-called 'variance components', i.e. the variation of the intercepts and slopes estimated for each individual are approximately Normally distributed with a mean of zero and some constant variance. The modelling procedure allows the testing of the variances to ensure they are non-zero.

As with the GEE and other regression-based approaches to longitudinal analysis, additional independent variables, at a categorical- or continuous-level, can be introduced to the model. Furthermore, it is possible to incorporate complicated interactions between time and the independent variables, including independent variables that vary over time. In the case of random coefficient models, not only can time intervals between measurements vary, but it is possible to have each individual measured at different times and with person-specific intervals between measurements. This makes this approach much more flexible, especially in situations where the interval between measurements is not controlled by the researcher, for example when studying consultation patterns.

13.3 Summary and discussion

The use of an interval-level measure as an outcome in epidemiological studies has the potential to address new and innovative research questions. However, it also raises some methodological issues that do not generally apply in analyses of the traditional epidemiological outcomes of onset and recovery. This chapter has outlined some of the possible approaches to

analysing longitudinal data that might be suitable for use in epidemiological studies. The choice of approach, or combination of approaches, will depend largely on the question that it is desired to answer.

The graphical approaches described at the start of this chapter provide a visual means by which to assess the distribution of an interval-level outcome and to asses changes in that outcome over time, and/or between population groups. For example, whilst an analytical method may inform as to the size of a change in the overall level of the outcome over time, graphical methods, such as Tukey mean-difference plots can show the reason for that change: a right shift in the distribution, or an increased number of people in the tail?

Methods suitable for only two times points, such as the paired t-test and the use of change scores, will be most appropriate when there is a particular interest in the difference in score between two time points, i.e. the interest is in the change. Although the GEE and random coefficients approaches are generally accepted as being more sophisticated and flexible methods of dealing with longitudinal data, they are in essence methods for the assessment of the association between independent variables, time varying or otherwise, and the 'shape' of the outcome over time.

In order to conduct a full and thorough analysis of the longitudinal development of an outcome and its association with other factors, it is likely that a combination of graphical and analytical methods will be required, considering both the absolute changes in the outcome over time and then how this is influenced by independent factors.

Having discussed these possible options to analyse interval-level longitudinal data in an epidemiological setting, the next chapter goes on to apply some of these methods to the new measure of locomotor disability developed in this thesis. In particular, the next chapter focuses on the potential of the interval-level measure to add information the field of locomotor disability research, over and above what is possible using a dichotomy of disability.

14 Modelling the longitudinal course of locomotor disability in the NorStOP1: changes over time and their association with socio-demographic factors and pain

14.1 Introduction

Chapters 9 to 12 described the development and testing of a new scoring mechanism for items from the PF-10 to create an interval-level measure of locomotor disability. This chapter illustrates some potential uses of this new measure in epidemiology, above and beyond those possible using a binary or ordinal definition of locomotor disability.

In particular, drawing on the methods described in Chapter 13, this chapter considers changes in the shape of the distribution of locomotor disability and how these relate to summary measures of change. These analyses investigate whether the groups defined by the socio-demographic factors and pain that have previously been shown to be associated with locomotor disability (see Chapters 3 and 8) have different patterns of change in locomotor disability over time.

In particular, this chapter considers the effect of socio-demographic factors and pain at follow-up in those pain-free at baseline in the NorStOP1.

14.2 Hypotheses

Box 14.1 defines the specific hypotheses around the course of locomotor disability to be tested in this chapter in those pain-free at baseline in the NorStOP1 subcohort.

The specific hypotheses around the association between pain and changes in locomotor disability in those pain-free at baseline are set out in Box 14.2.

Box 14.1	The course of locomotor disability in older adults pain-free at baseline

a. There is a right shift in the population distribution of locomotor disability over time, i.e. the average locomotor disability score in the cohort increases as the cohort gets older.

b. The rank of order of individuals, with respect to locomotor disability score, remains relatively stable over time despite the shift in the population distribution, i.e. those with relatively high locomotor disability scores at baseline continue to have relatively high scores throughout the follow-up period.

c. The rate of change in locomotor disability over time is progressive, i.e. the rate of change in locomotor disability score in one three-year period of follow-up is associated with the rate of change in a successive three-year period.

d. The rate of increase in locomotor disability over time will be higher in those who are older, those who are female, those who live alone and those in lower socioeconomic groups.

Box 14.2	The association between pain and changes in locomotor disability in older adults pain-free at baseline

e. There will be both a concurrent and a lagged effect of the onset of pain on the increase in locomotor disability, i.e. there will be a higher rate of increase in locomotor disability from baseline to three-year follow-up and from three to six-year follow-up in those who have an onset of pain from baseline to three-year follow-up than in those remaining pain-free at three-year follow-up. The rate of increase in both time periods will be highest in those who have an onset of pain in the lower limb and elsewhere.

f. In those who remain pain-free at three-year follow-up, there will be a higher rate of increase in locomotor disability from three- to six-year follow-up in those who have an onset of pain at six-year follow-up than in those remaining pain-free. The rate of increase will be highest in those who report an onset of pain in both the lower limb and elsewhere.

g. In those who have an onset of pain at three-year follow-up, there will be a lower rate of increase in locomotor disability from three- to six-year follow-up in those who recover from their pain at six-year follow-up than in those continuing to report pain. The rate of increase will be lowest in those who report an onset of pain in the lower limb and elsewhere and recover to have no pain.

14.3 Methods

14.3.1 Dataset

This chapter uses data from all three time points during the six years of the NorStOP1 subcohort (the only subcohort with six-year follow-up data within the time frame of this PhD). In the latter part of this chapter, the analyses examine the effect of changes in pain on the rates of change in locomotor disability. For this reason, this chapter uses data only from those people who were pain-free at baseline, i.e. responded “No” to the pain screening item and did not shade the body manikin (see Chapter 4 for details). Table 14.1 shows that 35% of those without pain at baseline were followed-up at both the three- and six-year stages of the NorStOP1. This is slightly higher than the follow-up rate of the subcohort as a whole (see Chapter 4). The gender structure and baseline mental health scores of the group pain-free at baseline and followed for six years was similar to those lost to follow-up during this time. However, those lost to follow-up tended to be older, belong to lower socioeconomic groups and have worse physical health at baseline than those who remained in the study, reflecting the follow-up of the NorStOP1 as a whole over this period (Chapter 4).

14.3.2 Outcome

The new measure of locomotor disability described in Chapters 10 to 12 was the outcome for use in these analyses. In order to fully utilise the Rasch measurement methodology across the three time points, i.e. baseline, three-year follow-up and six-year follow-up, a further Rasch analysis, as described in Chapter 9 was used. Scores were calculated using the data from the three time points together to ‘anchor’ scores over time and ensure that the ‘ruler’ did not change (Wright 2003).

14.3.3 Covariates

This section describes how the covariates of interest, i.e. socio-demographic factors and pain, were defined in the study and used in the analyses presented in this chapter.

Table 14.1 Follow-up of the NorStOP1 subcohort over six years: pain-free at baseline

n (%)	Followed-up to six-years	Not followed-up to six-years
Overall	680 (35)	1,265 (65)
Gender		
Male	321 (47)	591 (47)
Female	359 (53)	674 (53)
Age-group		
50 to 59 years	288 (42)	378 (30)
60 to 69 years	224 (33)	320 (25)
70 to 79 years	148 (22)	368 (29)
80 years and over	20 (3)	199 (16)
Educational attainment[a]		
Further education	128 (19)	138 (11)
School-age education only	544 (81)	1,099 (89)
Occupational class[a]		
Non-manual	292 (47)	333 (31)
Manual	333 (53)	746 (69)
Perceived adequacy of income[a]		
Adequate	479 (71)	732 (60)
Inadequate	198 (29)	496 (40)
Living arrangement[a]		
Not alone	521 (80)	1,596 (80)
Alone	132 (20)	403 (20)
SF-12 score at baseline[a]		
Physical component summary	51.4 (8.2)	47.3 (10.6)
Mental component summary	52.6 (8.9)	51.7 (9.6)

[a]Subject to missing data

14.3.3.1 Demographic characteristics and socioeconomic status

Gender, age, socioeconomic status and living arrangement were self-reported at baseline. Age was grouped into 10-year bands (50 to 59 years, 60 to 69 years, 70 years and over). In keeping with Chapter 8, three indicators of socioeconomic status were used: education (school-age only versus further education), occupational class (Office for National Statistics 2000 and 2002) (manual (lower supervisory/technical, semi-routine occupations, routine occupations) versus non-

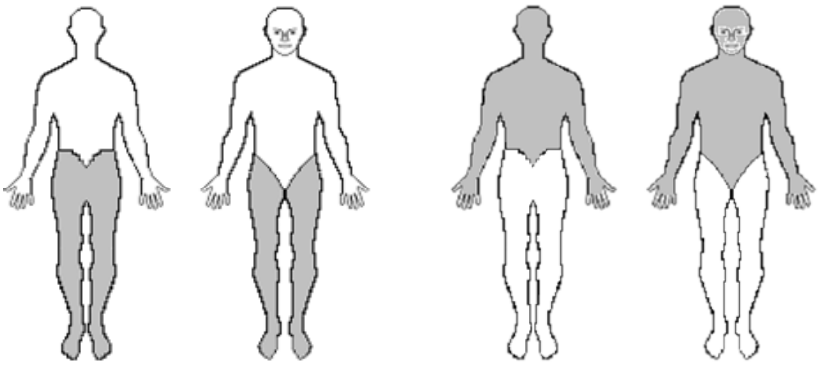
manual (higher managerial, higher professional, lower managerial/professional, intermediate occupations); self-employed people were excluded) and perceived adequacy of income (Thomas 1999) (inadequate ("*find it a strain to get by from week to week*", "*have to be careful with money*") versus inadequate ("*Able to manage without much difficulty*", "*quite comfortably off*")). Also considered was living arrangement, defined as alone or not alone. These indicators were described in detail in Chapter 4.

14.3.3.2 Presence and location of pain

Pain location was derived from the combination of the pain screening item and the body manikin, as described in Chapter 4. For this chapter, the specific pain areas examined are slightly different than in Chapter 8.

As this thesis is concerned with the measurement of lower limb locomotor disability, pain in the lower limb was thought to be of specific interest. Chapter 8 showed that pain in the low back had little additional impact on the cross-sectional association between lower limb pain and locomotor disability. This was in contrast to the stronger effect of pain in the lower limb and elsewhere (outside the lower limb and low back). For this reason, in the current chapter, pain in the low back (but not in the lower limb) was combined with pain 'elsewhere' hence, in the present analyses pain has been examined in two mutually exclusive areas: lower limb pain and elsewhere pain (Figure 14.1).

Figure 14.1 Definitions of lower limb pain and pain elsewhere



Lower limb pain Pain elsewhere

For the purpose of the analyses in this chapter, combinations of these pain areas are considered within a single, categorical variable: no pain (neither lower limb nor elsewhere), elsewhere pain only, lower limb pain only, and both lower limb and elsewhere pain. These data were collected at each of the three time points in the NorStOP.

14.3.4 Statistical analyses

14.3.4.1 Appropriate choice of analysis method

A series of options for the analysis of longitudinal interval-level data were discussed in the previous chapter, where it was highlighted that the most appropriate method to use would depend on the research question. Another major consideration was the presence of a large ceiling effect, and a smaller, although still noteable, floor effect in the locomotor disability score. This meant that the distribution of locomotor disability was highly skewed and that any assumption of Normality was unlikely to be met. This section considers each of the hypotheses in Box 14.1 and the most appropriate methods of analysis to investigate them.

Hypothesis a relates to the change in the distribution of locomotor disability over time. As discussed in Chapter 13, this is most effectively assessed using a graphical method, the most interpretable of which can be considered to be a Tukey mean-difference plot. In order to quantify any change in score, the calculation of a change score is required. This can be problematic in the presence of floor and ceiling effects. However, as it is the change itself that is of interest, the calculation of a change score, although limited in some respects, was deemed to be most appropriate method of analysis. Mean changes were calculated in each time period with associated 95% confidence intervals.

Hypotheses b and c relate to correlations between locomotor disability scores and changes in score over time. These hypotheses will therefore be assessed graphically using scatter plots.

Hypotheses d to g are concerned with rates of change in locomotor disability in specific time periods and how these vary between different groups. This means, as with Hypothesis a, that an approach to modelling change per se, rather than the shape of the outcome over time, is more appropriate. The presence of ceiling and floor effects not only affects the calculation of a change score, but also the method of analysing the association between changes in disability and other

factors of interest. Hence, although not necessarily the preferred method in most circumstances, linear regression analysis of the Normally distributed change score was the chosen method for these analyses.

14.3.4.1 Practical implementation of analyses

Tukey mean-difference plots

For the assessment of Hypothesis a, a Tukey-mean difference plot, as described in Chapter 13, was applied. This section describes how this plots was created and interpreted.

At each time point, locomotor disability scores were rank ordered and the sample split into groups of approximately 20 individuals per group (34 groups in the overall sample). The mean locomotor disability score in each of these groups was calculated. To compare the distribution of locomotor disability over time, the mean level of and difference (second time point minus the first) in disability in each rank-ordered group was calculated for each pair of time points. The mean (x-axis) and difference (y-axis) for each pair of rank-ordered groups were plotted on a scatter plot. Plots were weighted according to the number of rank-ordered groups that occupied the same coordinates. The plot was interpreted as described in Chapter 13.

Scatter plots

The scatter plots used to assess Hypotheses b and c were weighted by the number of observations at each set of coordinates. For Hypothesis c, a fitted regression line was added to the plot to aid interpretation.

Change scores

The analysis of Hypotheses a, d, e, f and g necessitated the calculation of change scores. Change scores were calculated as locomotor disability score at Time 2 (y_2) minus locomotor disability score at Time 1 (y_1). Thus a positive change score indicates that y_2 is greater than y_1, i.e. an increase in locomotor disability over time, and a negative change score indicates that y_1 is greater than y_2, i.e. a decrease in locomotor disability over time. Change scores can, in theory, range from –9.941 logits, i.e. change from most disability to most ability, to +9.941 logits, i.e.

change from most ability to most disability. Three possible change scores were considered as outcomes, i) baseline to three-year follow, ii) three-year to six-year follow-up, iii) baseline to six-year follow-up. To enhance comprehension for the reader, the meaning of the raw locomotor disability scores, the change in score and difference in change scores is laid out in Table 14.2.

Table 14.2 Summary of raw and change scores used in the analyses presented in this chapter

Lower limit	Upper limit
Raw score	
-4.906 **Least locomotor disability** as measured on scale	5.035 **Most locomotor** disability as measured on scale
Change score	
-9.941 **Improved:** changed from most locomotor disability to least locomotor disability	+9.941 **Deteriorated:** changed from least locomotor disability to most locomotor disability

Regression analysis

To investigate Hypotheses a, d, e, f and g, the change scores described above were entered into linear regression models as dependent variables. Independent variables of interest were gender, age-group, socioeconomic status (educational attainment, occupational class, perceived adequacy of income), living arrangement and pain status at three- and six-year follow-ups.

As discussed in Chapter 13, it has been shown (Glymour et al 2005) that the adjustment of change score models for baseline score leads to a 'double counting' of the score at the first time point and this can lead to inflated estimates of the regression coefficients. Hence the analyses of change presented in this chapter were not adjusted for the baseline locomotor disability score.

14.4 Results

14.4.1 Description of the sample

Approximately half of those without pain at baseline also reported themselves pain-free at three-year follow-up and at six-year follow-up (Table 14.3). The most common pain category at both three- and six-year follow-up was pain in both the lower limb and elsewhere. Overall, there was an increase in the level of locomotor disability in the sample over the six years of follow-up.

Table 14.3 Pain status and locomotor disability scores in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-up

	Baseline	Three-year follow-up	Six-year follow-up
Pain category[a,b], n (%)			
No pain	680 (100)	313 (47.6)	286 (44.6)
Lower limb pain only	-	94 (14.3)	82 (12.8)
Elsewhere pain only	-	65 (9.9)	63 (9.8)
Lower limb + elsewhere pain	-	186 (28.3)	210 (32.8)
Locomotor disability score (logits)[b,c], Median (IQR)	-4.91 (-4.91, -3.66)	-4.91 (-4.91, -3.66)	-3.66 (-4.91, -2.19)

[a]From body manikin (Figure 14.1); [b]Subject to missing data; [c]Higher score indicates more locomotor disability

14.4.2 The course of locomotor disability in older adults pain-free at baseline

Hypothesis a: *There is a right shift in the population distribution of locomotor disability over time*

There was a significant increase in the mean locomotor disability score between baseline and each of the follow-up time points (Table 14.4) in the NorStOP1 subcohort, and between the three- and six-year follow-ups.

Table 14.4 Change in locomotor disability score between time points in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=661)

	Change within groups: Mean (95% CI)
Baseline to three-year follow-up	0.52 (0.39, 0.66)
Three- to six-year follow-up	0.33 (0.18, 0.48)
Baseline to six-year follow-up	0.85 (0.69, 1.01)

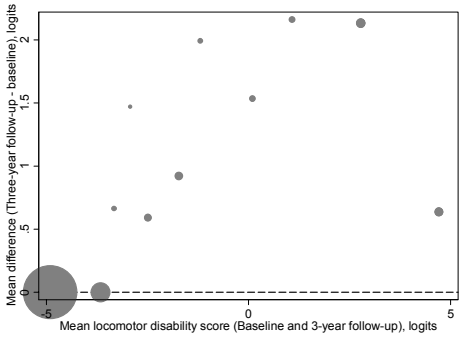
Positive change score indicates increase in locomotor disability

Figure 14.2 shows how these changes in mean came about as a result of the changes in the distribution of the locomotor disability between each pair of time points.

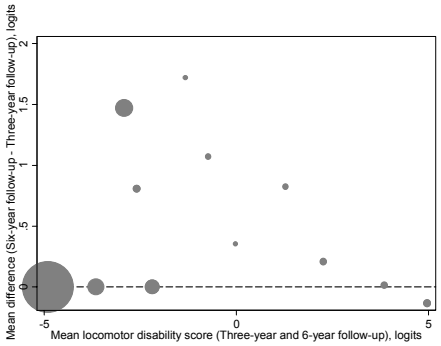
A very similar pattern is seen in the changes in the distribution of disability from baseline to three- and six-year follow-ups. From baseline to each of the follow-ups, there is a clear increase in the skewness of the distribution, and an overall increase in the level of locomotor disability (points lie above the zero line) (Figure 14.1(a) and (c)). Despite these changes in the distribution, those people who were in the ceiling of the distribution at baseline tended to remain in the ceiling or to display very little disability at the follow-ups. Those people with the highest mean scores over the baseline and follow-up time points tended to experience a decrease in locomotor disability over the follow-up periods (anomalous points on far right hand sides of Figures 14.1(a) and (c)). This is likely to be a result of the floor effect in the distribution of disability at baseline. From three-to six-year follow-up, a similar, but less clear, pattern was seen.

Figure 14.2 Change in the distribution of locomotor disability over time in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=661). Tukey mean-difference plots weighted to represent repeated values at the same co-ordinates

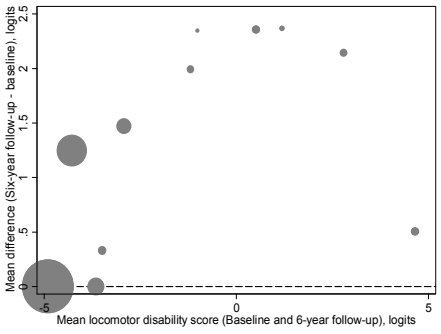
a. Baseline and three-year follow-up



b. Three and six-year follow-ups



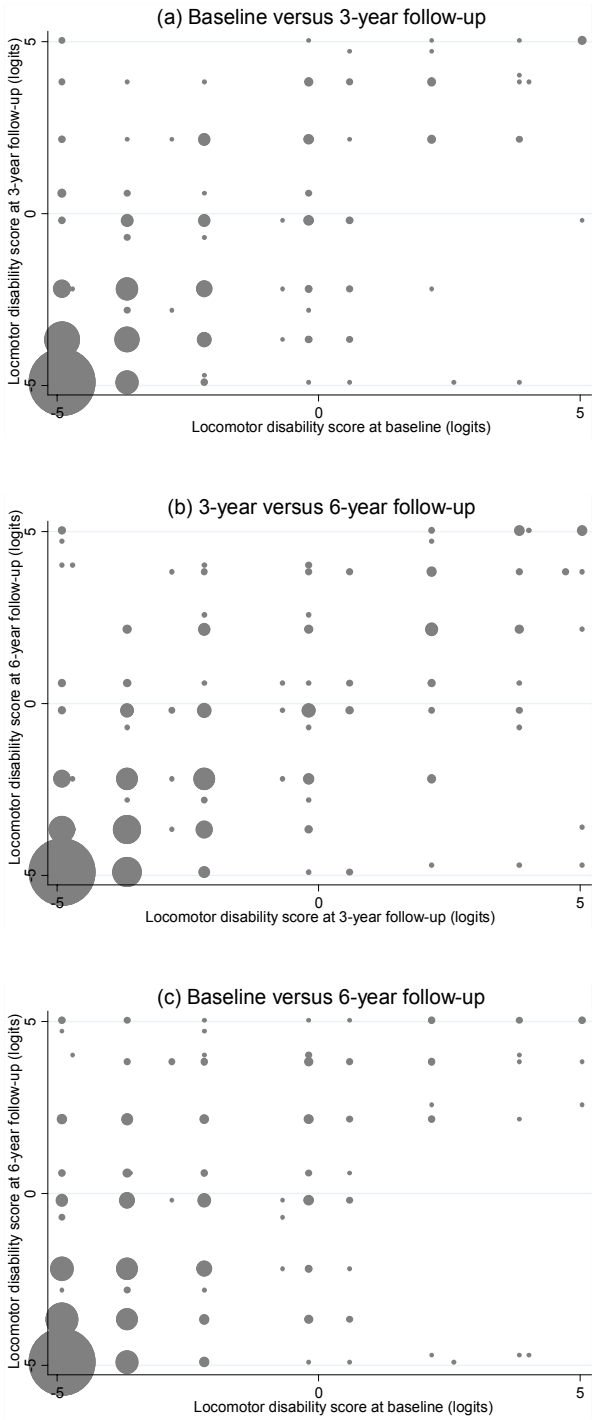
c. Baseline and six-year follow-up



Hypothesis b: *The rank of order of individuals remains relatively stable over time despite the shift in the population distribution*

Locomotor disability scores did not change between baseline and three-year follow-up and three- and six-year follow-ups for the majority of people (n=361 (55%) in each time period). Two hundred and sixty-six people (39%) had the same score at all three time points. Figure 14.3 compares the locomotor disability scores of individuals in between time points. Figure 14.3(a) shows that there is a positive association between scores at baseline and at three-year follow-up. There is a similar association between scores at three- and six-year follow-ups (Figure 14.3(b)) and across the whole six-year follow-up period (Figure 14.3(c)). This provides evidence in favour of Hypothesis b, that despite the right-shift in the population distribution, the rank ordering of people within the NorStOP1 subcohort remains relatively stable over time.

Figure 14.3 Locomotor disability scores at each wave of the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=661). Scatter plots weighted to represent repeated values at the same coordinates.



Hypothesis c: *The rate of change in locomotor disability over time is progressive*

Figure 14.4 shows that the majority of people experienced no change in locomotor disability score between baseline and three-year follow-up or between three- and six-year follow-ups in each tertile of baseline score. In those people who experienced an increase in locomotor disability score from baseline to three-year follow-up (right hand sides of Figure 14.4), the majority experienced a decrease in locomotor disability score from three- to six-year follow-up (lower right quadrants of Figures 14.4(a), (b) and (c)). In those people who experienced a decrease in locomotor disability score from baseline to three-year follow-up (left hand sides of Figures 14.4(a), (b) and (c)), the majority experienced an increase in locomotor disability score or remained unchanged from three- to six-year follow-up.

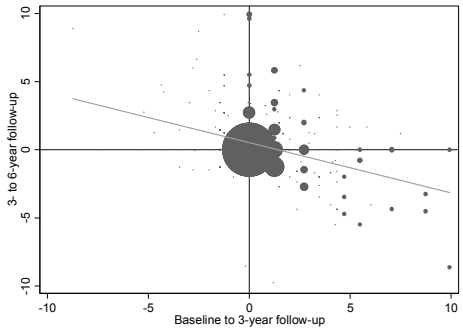
It was not possible for people the first tertile of the baseline locomotor disability score to experience a decrease in locomotor disability score from baseline to three-year follow-up as they were all in the ceiling of the score distribution at baseline. Within this group of people, there was a significant negative association between change scores in the two time periods (regression coefficient -0.42 (95% CI -0.52, -0.32)): an increase in locomotor disability score in the first period was generally associated with a decrease in score in the second (Figure 14.4(a)).

In those people with a score in the second tertile at baseline, a small minority increased their level of disability in both follow-up periods. As in the first tertile of the baseline score, there was a negative association (regression coefficient -0.25 (95% CI -0.48, -0.01)) between changes in the two time periods. Although this association was less marked than in the first tertile, on average, those with an increase in disability from baseline to three-year follow-up were likely to improve from three-to six-year follow-up. This tertile was less strongly influenced by the ceiling effect of the score at baseline than Tertile 1 (Figure 14.4(b)).

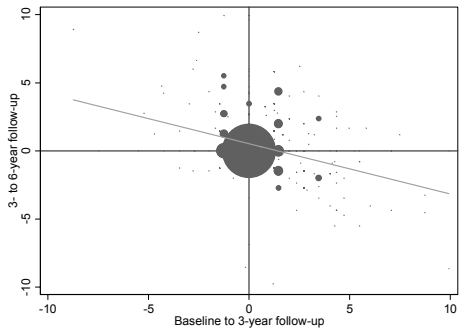
In people in the third tertile at baseline, there was also a negative association between changes scores across the time periods (regression coefficient -0.37 (95% CI -0.54, -0.20)). There was no floor effect within this tertile, hence this could not have affected the size of the change over time (Figure 14.4(c)).

Figure 14.4 Changes in locomotor disability score (logits) in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=661). Scatter plots weighted to represent repeated values at the same coordinates and displaying fitted regression line, by tertile of baseline locomotor disability score. Positive values represent faster rate of progression of locomotor disability

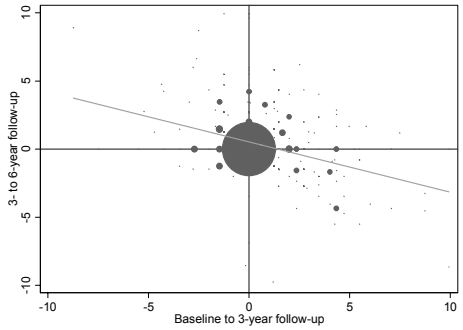
a. Baseline tertile 1



b. Baseline tertile 2



c. Baseline tertile 3



The large ceiling effect in the baseline locomotor disability scores, makes assessment of Hypothesis c less than simple. There is evidence to suggest that an increase in disability during the baseline to three-year follow-up period is associated with a decrease in disability between the three- and six-year follow-ups (overall regression coefficient -0.37 (95% CI -0.45, -0.29)). However, the effect was not so strong as to return people to their baseline level of locomotor disability at six-year follow-up, i.e. regression coefficient is greater than -1. This indicates that individuals did not fully recover from their acquired disability, and is in keeping with the findings in relation to Hypothesis a, in that there is a net increase in the level of disability in the cohort over time.

Hypothesis d: *The increase in locomotor disability over time will be larger in those who are older, those who are female, those who live alone and those with lower socioeconomic status*

Older age

Mean changes in locomotor disability score over time suggested a faster rate of progression of locomotor disability at older ages, which included a dose response association (Table 14.5).

Gender

Over the course of the six-year follow-up, average locomotor disability scores increased more rapidly in females than in males, in accordance with Hypothesis d (Table 14.6). However, the difference in mean change scores over time, between genders, was not significant.

Table 14.5 Change in locomotor disability score between time points in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=661). Mean change (95% CI) by age-group

		Age-group (years)		
		50 to 59 years	60 to 69 years	70 years and over
Change in locomotor disability score (logits)	Baseline to three-year follow-up	0.26 (0.05, 0.46)	0.53 (0.30, 0.46)	0.97 (0.70, 1.24)
	Between group difference [a], Mean (95% CI)	-	0.28 (-0.03, 0.58)	0.72 (0.38, 1.06)
	Three- to six-year follow-up	0.20 (-0.03, 0.43)	0.33 (0.07, 0.58)	0.57 (0.27, 0.87)
	Between group difference [a], Mean (95% CI)	-	0.13 (-0.22, 0.47)	0.37 (-0.01, 0.75)
	Baseline to six-year follow-up	0.46 (0.21, 0.71)	0.85 (0.57, 1.13)	1.52 (1.20, 1.85)
	Between group difference [a], Mean (95% CI)	-	0.39 (0.02, 0.77)	1.06 (0.66, 1.47)

Positive figures represent higher levels of disability and faster rate of progression of locomotor disability; [a]Difference in change score compared to 50 to 59 year age-group

Table 14.6 Change in locomotor disability score between time points in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=661). Mean change (95% CI) by gender in each follow-up period.

	Change within groups: Mean (95% CI)		Between group difference: Mean (95% CI)
	Female	Male	
Baseline to 3-year follow-up	0.59 (0.41, 0.78)	0.44 (0.24, 0.64)	0.16 (-0.11, 0.42)
3-6 year follow-up	0.33 (0.13, 0.54)	0.33 (0.11, 0.55)	0.00 (-0.29, 0.30)
Baseline to 6-year follow-up	0.94 (0.71, 1.16)	0.75 (0.51, 0.99)	0.19 (-0.14, 0.52)

Positive figures represent faster rate of progression of locomotor disability

Socioeconomic status

There was marginal evidence that those with school-age education only experienced a faster progression of locomotor disability than those with further education. This was the case for the baseline to three-year follow-up period only (Table 14.7). There was no further evidence of differences in the rate of progression between socioeconomic groups.

Table 14.7 Change in locomotor disability score between time points in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups. All values are Mean (95% CI)

	Within-group change in locomotor disability score (logits)		
	Baseline to three-year follow-up	Three- to six-year follow-up	Baseline to six-year follow-up
Educational attainment[a]			
School-age education only	0.58 (0.43, 0.73)	0.29 (0.12, 0.45)	0.87 (0.69, 1.05)
Further education	0.22 (-0.09, 0.54)	0.47 (0.12, 0.81)	0.69 (0.31, 1.06)
Between group difference in change score	0.36 (0.01, 0.70)	-0.18 (-0.56, 0.20)	0.18 (-0.24, 0.60)
Occupational class[a]			
Manual	0.61 (0.42, 0.81)	0.40 (0.19, 0.61)	0.80 (0.56, 1.05)
Non-manual	0.41 (0.20, 0.61)	0.38 (0.16, 0.61)	1.00 (0.77, 1.23)
Between group difference in change score	0.80 (-0.08, 0.49)	0.02 (-0.29, 0.32)	0.19 (-0.14, 0.53)
Perceived adequacy of income[a]			
Inadequate	0.51 (0.26, 0.76)	0.34 (0.06, 0.62)	0.84 (0.54, 1.14)
Adequate	0.53 (0.37, 0.69)	0.32 (0.15, 0.50)	0.86 (0.66, 1.05)
Between group difference in change score	-0.02 (-0.32, 0.28)	0.02 (-0.31, 0.35)	-0.02 (-0.38, 0.34)

Positive figures represent faster rate of progression of locomotor disability; [a]Subject to missing data

Living arrangement

Considering mean changes in the score over time, those who lived alone experienced a significantly higher rate of increase in disability than those who did not live alone from baseline to three-year follow-up (Table 14.8). This was not the case when considering the other follow-up periods.

Table 14.8 Change in locomotor disability score between time points in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=634). Mean change (95% CI) by living arrangement[a]

	Within-group change in locomotor disability score (logits)		
	Baseline to three-year follow-up	Three- to six-year follow-up	Baseline to six-year follow-up
Does not live alone	0.41 (0.25, 0.56)	0.35 (0.18, 0.52)	0.75 (0.57, 0.94)
Lives alone	0.89 (0.59, 1.19)	0.25 (-0.09, 0.58)	1.13 (0.77, 1.50)
Between group difference in change score	0.48 (0.15, 0.82)	-0.10 (-0.48, 0.28)	0.38 (-0.03, 0.80)

Positive figures represent faster rate of progression of locomotor disability; [a]subject to missing data

14.4.3 The association between pain and changes in locomotor disability in older adults pain-free at baseline

Hypothesis e: There will be both a concurrent and a lagged effect of the onset of pain on the increase in locomotor disability

Table 14.9 confirms the increased rate of progression of locomotor disability in those who developed pain at three-year follow-up compared to those remaining pain-free. When considering the site of this pain, only those with pain in lower limb and elsewhere had a significantly different mean change in disability than those without pain. However, all those with lower limb pain (alone or with elsewhere pain) experienced a significant increase in disability. This suggests a dose response association between the development of pain from baseline to three-year follow-up and concurrent change in locomotor disability score (Table 14.9).

Between three- and six-year follow-ups, those remaining pain-free at three-year follow-up had a faster rate of progression of locomotor disability than those who had developed pain at three-year follow-up (Table 14.9). This difference, which was not as a result of those with pain at three-year follow-up having acquired so much disability before the three-year follow-up so as not to be able to progress further in their disability (median (IQR) score at three-year follow-up: -3.659 (-4.906, -2.188)), was not statistically significant. A weak dose response association was seen between the onset of pain from baseline to three-year follow-up and change in locomotor disability from three- to six-year follow-up. Those with lower limb and elsewhere pain at three-year follow-up had a significantly smaller progression of disability than those with no pain.

When considering the whole six-year follow-up period of the NorStOP1, there was a significant increase in the level of disability over time, regardless of pain status at three-year follow-up. However, there was no significant difference in the change in locomotor disability score between pain groups. The only significant effect of pain on the rate of locomotor decline from baseline to six-year follow-up was in those people who experienced an onset of pain in both the lower limb and elsewhere from baseline to three-year follow-up.

The higher rate of increase in locomotor disability from baseline to three-year follow-up in those who experienced an onset of pain at three-year follow-up provided evidence to support Hypothesis e. Also as hypothesised, this rate of increase was highest in those with an onset of pain in the lower limb and elsewhere. However, there was no evidence to support the hypothesis that there was a lagged effect of this onset of pain between the three- and six-year follow-ups.

Table 14.9 Change in locomotor disability score between time points in the NorStOP1 subcohort: pain-free at baseline and completing three- and six-year follow-ups (n=640). Mean change (95% CI) by pain status at three-year follow-up

		Pain at three-year follow-up				
		No pain	Any pain	Site of pain		
				Elsewhere only	Lower limb only	Lower limb and elsewhere
Change in locomotor disability score (logits)	Baseline to three-year follow-up	0.27 (0.07, 0.47)	0.75 (0.56, 0.94)	0.23 (-0.13, 0.59)	0.65 (0.22, 1.08)	1.04 (0.79, 1.30)
	Between group difference[a,b], Mean (95% CI)	-	0.48 (0.21, 0.75)	-0.04 (-0.45, 0.37)	0.38 (-0.09, 0.85)	0.77 (0.45, 1.09)
	Three- to six-year follow-up	0.48 (0.26, 0.70)	0.19 (-0.12, 0.40)	0.31 (-0.10, 0.71)	0.31 (-0.17, 0.80)	0.10 (-0.19, 0.38)
	Between group difference[a,b], Mean (95% CI)	-	-0.29 (-0.59, 0.02)	-0.17 (-0.63, 0.29)	-0.17 (-0.70, 0.37)	-0.38 (-0.74, -0.02)
	Baseline to six-year follow-up	0.75 (0.51, 0.99)	0.95 (0.72, 1.18)	0.54 (0.10, 0.98)	0.95 (0.42, 1.48)	1.16 (0.85, 1.48)
	Between group difference[a,b], Mean (95% CI)	-	0.21 (-0.13, 0.54)	-0.20 (-0.71, 0.30)	0.20 (-0.38, 0.78)	0.42 (0.02, 0.81)

Positive figures represent a higher level of disability and a faster rate of progression; [a] Difference in change score compared to no pain group

Hypothesis f: *In those who remain pain-free at three-year follow-up, there will be a higher rate of increase in locomotor disability from three- to six-year follow-up in those who have an onset of pain at six-year follow-up than in those remaining pain-free*

In those people who were pain-free at baseline and three-year follow-up, there was a significant increase in locomotor disability score between the three- and six-year follow-ups, regardless of pain status at six-year follow-up. On average, this increase was significantly larger in those who experienced an onset of pain at six-year follow-up than in those who remained pain-free (Table 14.10). When considering the site of pain at six-year follow-up, the increase in disability score was only significantly larger than in the pain-free group in those having an onset of pain in lower limb and elsewhere.

Table 14.10 Change in locomotor disability score between three- and six-year follow-ups in the NorStOP1 subcohort: pain-free at baseline and three-year follow-up (n=288). Mean change (95% CI) by pain status at six-year follow-up.

		Mean (95%CI) change in locomotor disability score (logits)	Between group difference, Mean (95% CI) (logits)[b]
Pain status at six-year follow-up[a]	No pain	0.28 (0.04, 0.53)	-
	Any pain	0.73 (0.44, 1.01)	0.45 (0.07, 0.82)
	Elsewhere only	0.57 (0.01, 1.13)	0.28 (-0.33, 0.89)
	Lower limb only	0.70 (0.01, 1.39)	0.42 (-0.32, 1.15)
	Lower limb and elsewhere	0.81 (0.43, 1.19)	0.53 (0.08, 0.98)

Positive figures represent faster rate of progression of locomotor disability; [a]Subject to missing data; [b]Relative to no pain group

There is evidence in favour of Hypothesis f in that those people pain-free at baseline and three-year follow-up who experience and onset of pain at six-year follow-up had a faster rate of increase in locomotor disability between three- and six-year follow-ups and this increase was largest in those with an onset of pain in the lower limb and elsewhere. This finding strengthens the conclusions from Hypothesis e that an onset of pain over a three-year period is associated with a concurrent increase in locomotor disability.

Hypothesis g: *In those who have an onset of pain at three-year follow-up, there will be a lower rate of increase in locomotor disability from three- to six-year follow-up in those who recover from their pain at six-year follow-up than in those continuing to report pain*

Those people who reported an onset of pain at three-year follow-up and continued to have pain at six-year follow-up, on average, experienced a statistically significant progression in their disability between three- and six-year follow-ups (Table 14.11). This was not the case for those who had recovered from their pain at the six-year follow-up. Those recovering from their pain at six-year follow-up, on average, experience a decrease in locomotor disability.

In those with single site pain at three-year follow-up, there were no significant differences in levels of change in locomotor disability score between the pain groups at six-year follow-up. In those with both lower limb and elsewhere pain at three-year follow-up, those returning to no pain at six-year follow-up experienced a decrease in their level of disability from three- to six-year follow-up, although the change within this group was not statistically significant. However, the change in disability was significantly greater than that experienced by those continuing to have both lower limb and elsewhere pain at six-year follow-up.

Table 14.11 Change in locomotor disability score between three- and six-year follow-ups in the NorStOP1 subcohort: pain-free at baseline and onset of pain at three-year follow-up (n=345). Mean change (95% CI) by pain status at six-year follow-up.

		Pain at six-year follow-up				
		No pain	Any pain	Site of pain		
				Elsewhere only	Lower limb only	Lower limb and elsewhere
Change in locomotor disability score (logits)	Any pain at three-year follow-up[a] (n=324)	-0.12 (-0.54, 0.30)	0.34 (0.06, 0.62)	0.01 (-0.60, 0.62)	0.63 (-0.04, 1.30)	0.37 (0.01, 0.74)
	Between group difference, Mean (95% CI)	-	0.46 (-0.05, 0.97)	0.13 (-0.62, 0.87)	0.75 (-0.04, 1.54)	0.49 (-0.07, 1.05)
	Single site pain at three-year follow-up[a] (n=148)	0.22 (-0.33, 0.77)	0.36 (-0.05, 0.77)	0.02 (-0.71, 0.74)	0.93 (0.15, 0.74)	0.93 (0.15, 1.71)
	Between group difference, Mean (95% CI)	-	0.14 (-0.55, 0.82)	-0.20 (-1.11, 0.71)	0.71 (-0.24, 1.67)	0.02 (-0.82, 0.86)
	Lower limb pain and pain elsewhere at three-year follow-up[a] (n=176)	-0.48 (-1.13, 0.16)	0.33 (-0.06, 0.72)	-0.00 (-1.06, 1.05)	0.07 (-1.12, 1.27)	0.43 (-0.03, 0.88)
	Between group difference, Mean (95% CI)	-	0.81 (0.06, 1.57)	0.48 (-0.75, 1.72)	0.56 (-0.80, 1.92)	0.91 (0.12, 1.70)

Positive figures represent faster rate of progression of locomotor disability; Single site pain: lower limb pain only or pain elsewhere only; [a]Subject to missing data

14.5 Discussion

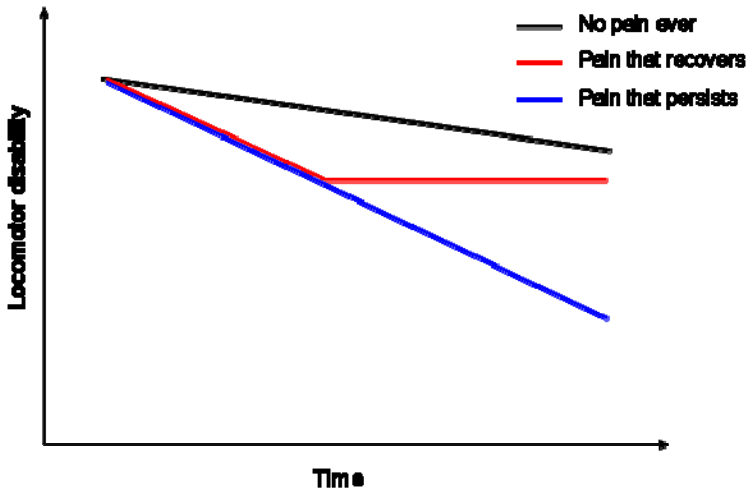
14.5.1 Principal findings

This chapter has presented analyses concerning the locomotor disability of those people in the NorStOP1 subcohort who were pain-free at baseline and who responded to the surveys at baseline, three-year and six-year follow-ups. Analyses were conducted using the interval-level measure of locomotor disability developed this thesis.

The analyses presented in this chapter have provided evidence that there is a shift in the level of disability in the population over time, as people get older, and that as this shift occurs, individuals tend to retain their rank ordering in the population. However, it was not the case that the rate of change in locomotor disability was progressive. Indeed, rather than an increase in disability in one time period being indicative of a future increase in disability, it was, on average, followed by a decrease in the level of disability. Despite this decrease in disability after an increase, on average, those experiencing an increase followed by a decrease in disability did not return to their baseline level of disability in the second period, i.e. recovery was not complete. There was no association between gender or socioeconomic status (educational attainment, occupational class, perceived adequacy of income) and the rate of change in locomotor disability. Living alone at baseline was associated with a higher rate of increase in disability, as was older age, with a dose response association being seen across 10-year age-groups.

The onset of pain was associated with a concurrent increase in locomotor disability. This association was strongest in those with an onset of pain in the lower limb and elsewhere. There was a non-significant lagged association between the onset of pain and the rate of change in locomotor disability. Recovery following the onset of pain resulted in a decrease in the level of locomotor disability, relative to those who continued to have pain, but did not return these people to the level of functioning experienced by those who never experienced pain (Figure 14.5). Over time, this would created a cumulative effect, with those developing and recovering from pain having higher levels of disability than those never experiencing pain.

Figure 14.5 Schematic representation of the effect of pain on levels of locomotor disability over time



14.5.2 Strengths and weaknesses of the study

The use of an interval-level measure of locomotor disability in these analyses has allowed some new and innovative questions to be asked about the association between pain and locomotor disability. However, it has also raised some methodological issues that do not generally apply in analyses of the onset of or recovery from disability.

First, the interval-level measure has allowed the shape of the distribution of disability and changes in this shape in the population to be considered, as suggested necessary by Rose (1992) in a general context, and Thomas (2007) in the more specific context of disability. Furthermore, the level of disability at a particular time has been assessed as a predictor of the level of disability in the future. This is more informative than looking only the onset and persistence of disability, because it gives a fuller picture of the situation.

Second, having a true measure of disability allows a fine-grained approach to disability measurement. Even if one does not consider the distribution of disability in the population as a whole, one can more accurately assess a person's level of disability, rather than assigning them simply to a state of ability or disability. Although for some purposes, for example allocation of state benefits, it is necessary to create a dichotomy, this is not necessarily helpful for research or clinical purposes.

Third, the measure allows several possible methods to be used in the study of change in score over time. In the case of the analyses presented here, the interest was in change in score between two time points for example baseline to three-year follow-up. The decision was taken in this chapter to calculate a difference in score over time and use this as the outcome for the analyses. An alternative to this approach would have been to model the absolute level of locomotor disability at the end of the follow-up period, for example three-year follow-up, and to adjust the analysis for baseline locomotor disability score. This is a mathematically similar approach that has been said to reduce the problem of regression to the mean, but was problematic in the NorStOP1 dataset because of the skewed distribution of the locomotor disability score.

The presence of both floor and ceiling effects in the locomotor disability score in the NorStOP1 represents a limitation of these analyses. Most notably, change scores for those people with scores in the ceiling are not accurate, because they are censored, i.e. it is not possible to know how far someone has moved along the latent trait.

Previous analyses in this thesis have shown a strong association between pain and the presence of locomotor disability cross-sectionally at baseline in the NorStOP (Chapter 8). In this chapter, having chosen a group of individuals who were pain-free, the group being analysed may have had lower levels of locomotor disability at baseline than was usual for them. Hence on re-measurement, they would be most likely to move closer to their long term, average level of disability, i.e. to appear more disabled. This is a statistical phenomenon known as regression to the mean (Barnett et al 2005). The ceiling effect in the locomotor disability score is likely to have exacerbated this effect, because in people with low levels of disability, who scored in the ceiling at baseline, it was only possible to see increases in their level of disability. Regression to the mean may therefore be partly responsible for the general tendency of people who are pain-free at baseline to experience an increase in disability over time (as shown in Figure 14.2). However, this seems unlikely to be the prevailing reason when one considers the ageing of the population and the changes in disability in relation to changes in pain status, which are broadly analogous to the literature relating to the association between prevalent pain at baseline and the future onset of locomotor disability (Clark et al 1998a and 1998b, Ebrahim et al 2000, Guralnik et al 1993, Leveille et al 2007, Wannamethee et al 2005). Furthermore, this group of people who were pain-free at

baseline are an unusual group in the older adult population, where the majority of people report at least some pain in a one-month period (Thomas et al 2004b).

The association between change in pain and change in locomotor disability was shown to be contemporaneous, rather than the pain becoming present before the change in disability. However, this could be as a result of the three-year follow-up intervals, which may also account for the lack of finding in the investigation of the lagged effect of changes in pain on changes in disability. This assessment of pain in the last month at three-yearly intervals is a major limitation in these analyses of the association between pain and changes in locomotor disability. This method of pain recording clearly misses the dynamic nature of pain (Gureje et al 2001, Dunn et al 2006), and will underestimate the prevalence of pain in the population. It is possible that the lagged effect of pain on locomotor disability is shorter than three years. If so, any lagged association would likely be diluted in those who developed pain soon after baseline because the 'lagged' change in locomotor disability could have occurred before three-year follow-up; thus appearing to be a concurrent change. A further limitation of analyses into the lagged effect of pain is that although these people were all pain-free at baseline, they had developed pain at different locations and at different times before the three-year follow-up. Assuming that pain has an effect on the level of locomotor disability concurrently within a three-year period, these people will start the three- to six-year follow-up period with differing durations of pain.

14.5.3 Strengths and weaknesses of the study in relation to the current literature

The findings from this study relating to the association between gender and socioeconomic status and changes in disability may seem to be at odds with the current literature on the onset of locomotor disability. However, the analyses of factors associated with onset in the current literature are addressing a slightly different question to the analyses presented in this chapter. Although these differences may be due to the pain-free nature of the current sample at baseline, there is no reason to suppose that, for example, because pain at baseline is associated with the onset of locomotor disability (Leveille et al 2007), that individuals with an onset of pain should have a faster rate of increase in disability than those remaining pain-free. This distinction in question is important,

and clearly illustrates the additional information that can be gained through the use of this new measure of locomotor disability.

Previous studies of the association between pain and the onset of locomotor disability have considered only prevalent pain at baseline. Therefore, the analyses presented in this chapter that considered the association between pain and locomotor disability, add to the literature in two ways. Not only do the analyses presented here consider the onset of and recovery from pain, allowing at least in part for its well-known dynamic nature (Gureje et al 2001, Dunn et al 2006), they also consider changes in the extent of locomotor disability over time, rather than simply its presence or absence.

The analyses in Chapter 8 were adjusted for the potentially confounding factors of age, gender, socioeconomic status and living arrangement. This was not the case in this chapter, as associations were assessed in a more qualitative manner. There is therefore the potential for confounding in the association between changes in pain and changes in locomotor disability. In addition to the potential confounders adjusted for in Chapter 8, previous studies have often made adjustments for the presence of comorbid disease including cardiovascular disease, cancer and stroke (for example Clark et al (1998a), Leveille et al (2007), Wannamethee et al (2005)) and lifestyle factors such as smoking, BMI and activity levels (for example Clark et al (1998a) Ebrahim et al (2000)). However, these adjustments have not generally attenuated the pain-disability association. Although, the analyses presented in this chapter relate specifically to changes in the level of disability and their association with changes in pain, which as discussed above is different to previous studies presented in the literature, the continued association in the literature between pain and disability after adjustment for other confounders suggests that these factors do not substantially alter such associations and that adjustment for them here would not have substantially altered the findings presented in this chapter.

14.5.4 Meaning of the study

This chapter has shown that further investigation of the association between pain and changes in locomotor disability is possible using an interval-level measure of disability compared to using a simple binary definition of disability. In particular, it has allowed the investigation of the

distribution of locomotor disability over time and between groups, as well as the modelling of the extent of disability over time in relation to other factors, notably pain. The use of this measure of locomotor disability in these analyses has provided additional information regarding the association between changes in pain and changes in locomotor disability that has previously been unavailable due to the lack of such a measure.

14.5.5 Unanswered questions and future research

An important caveat of the analyses presented in this chapter is that they relate to a select group of people aged 50 years and over who were free of pain at baseline. This is unusual in a group of community dwelling older adults, where the level of reported pain has been shown to be high (Thomas et al 2004b). It remains to further investigate the association between changes in pain status and changes in disability within those people with prevalent pain.

What the analyses presented in this chapter have been unable to determine is the clinical implication of a change in pain status on a change in locomotor disability. The reasons for this are two-fold. First, the analyses presented in this chapter did not take account of the intensity or duration of the pain that was reported. It could be that the extent of the spread of pain is confounded by the intensity of pain, and so to conclude changes in disability are purely as a result of the pattern of pain, rather than other features of the pain would be too strong. Further analyses are needed, accounting for the nature and duration of pain.

Second, as mentioned above, no minimally important change has yet been calculated for the locomotor disability score and so the sizes of the changes in disability over time are difficult to interpret. The development of such a clinically important change should be a priority for future research to enable a more efficient and useful implementation of this new measure of locomotor disability and more useful interpretations of associations, such as those presented in this chapter. In order to establish such a clinically important change value, additional data would be required, including some form of ‘anchor’ with which to define the size of the change. As discussed in Chapter 12, such an anchor was not available in the NorStOP datasets.

To date, the lack of a measure of locomotor disability with the appropriate mathematical properties has hampered a detailed investigation of the course of locomotor disability over time.

Previous researchers in this area have been able to look at gross changes in locomotor disability, but have been unable to look at finer changes in disability or the quantitative differences between levels of disability. For example, Guralnik et al (2001) defined levels of walking disability (none – able to walk half a mile without help, moderate – unable to walk half a mile without help, severe – help needed or unable to walk across a small room). These authors considered progress through these levels, in terms of progressive and catastrophic onset of severe disability. It has been possible for researchers to investigate trajectories using binary definitions of disability by employing techniques such as latent class analysis (Clogg 1995). However, with binary definitions, latent class analysis only allows a person to be assigned to a category, for example disabled or not at each time point, and to group people according to these categories over time. A potential use of an interval-level measure of locomotor disability is to look at the course of disability in more detail, either on an individual or a group level, say in the context of a random coefficient model. Furthermore, the use of binary and ordinal scale definitions does not allow for the quantification of changes over time. Analysis of this type was restricted in the NorStOP1, as there were only three time points, but with more repeated measurements, such an analysis would be helpful in the assessment of locomotor disability.

Furthermore, it remains to be considered whether locomotor disability acts as a continuous risk factor for other disabilities, such as difficulties bathing, or feeding, (Dunlop et al 1997), or through social isolation (Iezzoni et al 2001). As described by Rose (1985) in relation to the association between the distribution of blood pressure and cardiovascular disease, this could be potentially very important. If many people have a low level of locomotor disability, then targeting interventions at the population may reduce the overall level of disability and so reduce its effect in causing the other disabilities for which it is a risk factor. Similarly to Rose's discussion of a binary cut off to define hypertension, the use of a binary definition of locomotor disability may not allow full consideration of the problem and the most effective way to combat it. The analyses presented here have shown, in keeping with the theory of Rose (1985), that the tail of the locomotor disability distribution becomes heavier over time and despite a non-Normal score distribution, this is associated with an increased mean level of locomotor disability.

14.6 Conclusions

This chapter has illustrated the ability of an interval-level measure to allow a detailed analysis of the distribution and course of locomotor disability and its association with selected risk factors. In the course of this illustration, the association between changes in pain and concurrent increases in locomotor disability, above and beyond the background increase over time, have been highlighted.

The next chapter reviews the position at the beginning of this PhD project and considers the progress towards achieving the aims of this thesis. This focuses on the development of the new measure of locomotor disability, and its potential usefulness in future research in this field.

15 Discussion and conclusions

15.1 Introduction

This thesis arose out of the convergence of two key areas: i) the epidemiology of locomotor disability in middle- and old-age and its relationship to musculoskeletal pain, and ii) the concept of disability as a continuum rather than a discrete state. This chapter presents a brief review of the work undertaken in this thesis and its main findings followed by a critical reflection on the methods used and the contribution of this work to knowledge in this field. The chapter concludes by considering the implications of this thesis for future research and practice.

15.2 Principal findings

This section describes the key findings presented in this thesis, a brief summary of which is provided in Box 15.1.

A large array of terms has been used to describe locomotor disability and these characterisations centred on binary definitions of disability that varied greatly between studies. The lack of agreed definition of disability made comparison across studies difficult, whilst the lack of an interval-level measure did not mirror the natural occurrence of locomotor disability.

The review of the literature did highlight the use, by some authors, of the widely employed PF-10 in the assessment of locomotor disability (Bohannon et al 2004, Peat et al 2006a, Syddall et al 2009). Comparison of the item content to the ICF showed that the PF-10 covered too wide a range of tasks to assess only the construct of locomotor disability. Furthermore, the summated-rating scoring mechanism did not result in an interval-level of measurement.

Five individual items from the PF-10 did however map well to aspects of the ICF related to locomotor disability, as it was defined in this thesis. The psychometric properties of these items, including repeatability, construct validity and feasibility were tested and confirmed in the NorStOP datasets. In particular, these items are already so widely used (Garratt et al 2002) that their application to assess locomotor disability presents little extra burden to either researchers or participants.

To explore the potential of these individual items in study the epidemiology of locomotor disability, ordinal regression models were used to assess the cross-sectional association between socio-demographic factors and pain. Strong associations were seen between locomotor disability and older age. Also evident was an association between prevalent pain and locomotor disability, which was noticeably stronger in middle- than in old-age.

Despite the potential of these ordinal-level item responses to provide more detailed assessments of locomotor disability than binary definitions, they did not offer interval-level measurement. The Rasch model was therefore used to create an interval-level measure of locomotor disability from the five PF-10 locomotor disability-specific items. Due to local response dependency between the items, this required the formation of two super-items: one for walking and one for stair-climbing. The generalisability of the super-item scoring mechanism was established in datasets external to that used to derive the mechanism, and the psychometric properties of unidimensionality, repeatability, construct validity, responsiveness and interpretation were tested and confirmed. A substantial ceiling effect was detected in both genders up to the age of 70 years and a floor effect in those aged 80 years and over.

Box 15.1 Key findings from the thesis

- There is a lack of consensus on the conceptual definition and operational measurement of locomotor disability in population studies, with no use of interval-level measures
- PF-10 subscale score is a candidate measure of locomotor disability, but item content is too wide and the scoring mechanism does not provide interval-level measurement
- Individual items of the PF-10 provide psychometrically sound, ordinal-level indicators of the level of locomotor disability
- Ordinal regression models showed a cross-sectional association between socio-demographic factors, pain and locomotor disability, providing additional information about these associations than a simple binary definition of locomotor disability
- The Rasch model specification was used to derive a generalisable, repeatable and valid interval-level measure of locomotor disability from the five locomotor disability-specific items of the PF-10 through the formation of super-items
- The interval-level measure of locomotor disability allows the investigation of the distribution of disability and in particular use of the distribution to explain mean changes in the level of disability in a population

Potential uses of the new measure of locomotor disability were illustrated by considering the distribution of locomotor disability over time in those pain-free at baseline. Locomotor disability was found, on average to increase over time, as the sample aged. This increase was faster in those people with an onset of pain. Recovery from pain was accompanied by a relative decrease in disability, but these people were not returned to the same level of disability as had the pain never been present.

15.3 Key decisions and their implications for the interpretation of this thesis

15.3.1 The exclusive use of items from the PF-10

A key decision in the course of this thesis was to restrict the items for the new measure of locomotor disability to those from the PF-10. Whilst the PF-10 has been suggested as a measure of locomotor disability (Syddall et al 2009), and previously been used in this capacity (Peat et al 2006a), this approach limited the items available for inclusion in the measure. Having removed from the potential item set those items that did not map to concepts of the ICF that fell within the definition of locomotor disability adopted in this thesis, only five items remained. The content of these items was very homogenous, covering only walking and stair-climbing tasks; meaning that the content validity of the measure was somewhat limited, with the measure covering only the absolute core tasks generally considered to be related to locomotor disability (see Table 3.1). Furthermore, each of the five items included had only three response options. This resulted in coverage of only a section of the continuum of locomotor disability, and consequently the presence of ceiling and floor effects in the disability distribution in the NorStOP. This issue is discussed further in Section 15.3.4.

Despite these limitations, the PF-10 was carried forward as the source of items for the new measure of locomotor disability. This decision was taken for several reasons. First, although the items do not allow the coverage of the whole spectrum of disability, they allow those with levels of functioning below the level of ‘no limitation in walking more than a mile and climbing several flights of stairs’ and above ‘a lot of limitation in walking 100 yards and climbing one flight of stairs’ to be identified. These levels could reasonably represent the two ends of the spectrum considered

by Thomas (2007), when he stated that those at the ends can be identified easily, whilst those in the middle require more differentiation. For example, those in the floor of the distribution would likely be unable to carryout basic daily tasks such as walking to the bus stop or local shop, whilst those in the ceiling of the distribution would be able to walk around a large supermarket or walk for pleasure without any difficulty. These could be considered clinically meaningful levels of disability: it would be feasible for the desirable outcome from an intervention in research or in practice to be the ability to walk 100 yards or climb a single flight of stairs. Similarly, a public health outcome of interest might be for a particular proportion of the population to be able to walk more than a mile or climb several flights of stairs without limitation. Whilst this might be reasonable in a general population, in a lower functioning population such as the oldest old or a patient population, the restriction to the items of the PF-10 could result in the presence of a larger floor effect. This would present similar problems to those seen in the NorStOP population in terms of the skew of the score.

This is not to say however, that should it have been possible to create new items for this measure of locomotor disability that these levels of disability would have been chosen as the ends of the scale. This approach though would have meant the development of another health assessment instrument to add to the thousands already available to researchers and clinicians (Feinstein 1986, de Vet et al 2006). This would have been costly in terms of time and resources, and it is not guaranteed that this would have led to an instrument with improved psychometric properties.

Despite some limitations in the content of the new locomotor disability measure, it does add to the breath of the continuum of disability that can be measured by most performance-based assessments, in which ‘easier’ tasks, such as walking shorter distances are usually addressed. It also brings in some elements of person-perceived limitations that are not assessed by performance measures.

To abandon the PF-10 in favour of a de novo measure for locomotor disability would have been to ignore the most widely used self-reported health status assessment instrument in the world (Garratt et al 2002). Other authors have used selections of items from across previously published measures. For example, in the development of the LEFS Binkley et al (1999) reviewed existing

questionnaire items and interviewed clinicians, whilst the recent PROMIS initiative (Fries et al 2009) has collated and reviewed items from many pre-existing scales, including the SF-36, with a view to their inclusion in an item bank. However, the use of items from more than one existing scale, or the development of new items, means that data in which to develop and test the new measure would not be as easily available, if at all. This would slow the instrument development process, and potentially add to patient burden if more items needed to be included in each study wishing to use the measure. This could lead to a smaller uptake of the use of the scale by researchers and lower response rates where the scale is used.

Recent work by several groups has suggested the use of item banking and computer adaptive testing (CAT). In particular, the PROMIS initiative for example, is taking items from a wide range of sources, including the SF-36, and assessing them using item response theory (IRT) in order to compile a comprehensive set of items for various health-related concepts including physical functioning. Whilst this approach will provide good content validity and a much wider range of content coverage than the approach adopted in this thesis, it will not produce an interval-level measure of locomotor disability because, as mentioned previously, only the Rasch model can produce true measurement: other forms of IRT models do not do this (see below). Furthermore, though the use of CAT can produce more precise estimates of the level of disability and potentially remove ceiling and floor effects from the locomotor disability distribution, it cannot be used in postal surveys: a computer is needed. The main advantage of CAT though is its ability to reduce participant burden, and in the case of the locomotor disability measure, this is minimal, as there are only five items.

15.3.2 The effect of local response dependency and the use of super-items

The Rasch model (Rasch 1960) was chosen to create the measure of locomotor disability because, mathematically, it is the only method by which ordinal item responses can be transformed to an interval-level measure (Fischer 1995). Item response theory (IRT) models, and in particular the one-parameter IRT model, which is mathematically equivalent to the Rasch model, are sometimes seen as equivalent to this approach. However, they originate from a different paradigm: that of modifying the model to create a good fit to the data (Andrich 2004). This paradigm is not

aimed at the creation of an interval-level measure, and so the Rasch (one-parameter IRT) model was the only available approach.

On full investigation of the theory of the Rasch model specification, it became clear that despite their acceptable psychometric properties, the raw PF-10 items were not suitable for combination into a measure using Rasch analysis due to the local response dependency between the items. At this point, it was considered whether it was worthwhile pursuing the development of an interval-level measure using these items. The widespread use and availability of the PF-10 to a wide range of researchers made the idea of abandoning the PF-10 unattractive and lead to the consideration of methods to manipulate these item responses to fit the Rasch model specification. The RUMM2020 software used throughout this thesis has an inbuilt 'subtest' function to remove local response dependency from the set of items. However, in the case of the PF-10 data in the NorStOP pilot dataset, this did not solve the problem of overall misfit to the specification. A more considered approach to the problem was required. The items clearly formed two hierarchies (walking and stair-climbing), with several logical orderings of the item thresholds possible. Empirical orderings of the item thresholds were assessed, and the most frequently occurring pattern used in creating two super-items. This still did not lead to acceptable fit to the Rasch model, as the thresholds for the super-items were not in the order expected by the model. In this situation with raw items, the standard approach to this problem is to combine categories of the item responses to produce ordered thresholds. However, this is not standard practice in the case of the subtests created by RUMM2020, because the categories of the subtests cannot be interpreted. In the case of the super-items, categories had been created by considering the relative responses to items, and so it was considered acceptable to combine categories. Indeed, this actually provided additional information about what levels of limitation in different difficulties of task could be considered equivalent. For example, given that a report of "a little limitation" in climbing one flight of stairs, it is not possible to distinguish the level of locomotor disability between individuals based on reports of "a little limitation" and "a lot of limitation" in climbing several flights of stairs (Table 10.3).

The use of super-items, followed by the combination of categories of these super-items to derive an interval-level measure from a series of the raw ordinal-level item responses is somewhat

unorthodox and not commonly used. This begs the question of whether the approach is suitable, or whether the measure is too far removed from the original item responses.

A substantial amount of data manipulation is required in order to calculate a score on the locomotor disability measure from the raw items. However, it is possible to relate the scale scores back to responses to the raw items (Figure 10.5). This is not possible in many currently used scores created by summated ratings, for example the original PF-10 scale (Ware & Sherbourne 1992). Thomas (2007) raised this point of the usefulness of scales such as the SF-36 because they assign values to people in terms of their level of disability, but these values seldom have a qualitative meaning: they merely allow people to be compared to each other. The new measure of locomotor disability developed in this thesis, despite its complexity of scoring, allows this qualitative meaning to be attached. Even with this qualitative meaning, a major limitation of the new measure, which stems partly from the restriction to the items of the PF-10, is the complex scoring mechanism. Whilst it would not be straightforward to create a score from any items using Rasch analysis, the use of the super-items makes this particularly difficult, requiring specialist software and expertise. This may limit the uptake of the measure in practice.

Another question arising from the amount of data manipulation undertaken to create the locomotor disability measure could be whether the fit of the super-items to the Rasch specification is reproducible. The widespread use of the PF-10 allowed this issue to be addressed. The scoring mechanism was robust in three datasets representing local, national and international comparisons. Whilst this cannot ensure the fit of the super-items to the Rasch model in any dataset, fit of the data to the specification in these three datasets is reassuring and suggests that the super-item scoring mechanism is not specific to the NorStOP pilot dataset.

15.3.3 The presence of floor and ceiling effects in the new measure

A major criticism of the new measure of locomotor disability has to be the presence of floor and ceiling effects (Table 12.2). These effects arose due to the measure's limited coverage of the continuum of locomotor disability, as discussed above.

Changes in score between two time points are heavily influenced by the presence of such effects. This is because it is not possible to accurately measure real change in score in people who

have a score at one or both time points in the ceiling or the floor of the distribution. For those people with an extreme locomotor disability score at both baseline and follow-up, it is not clear whether their actual level of locomotor disability remained the same over time, or whether they experienced changes that could not be detected by the measure. For those people with a score at baseline or follow-up in the ceiling or floor of the distribution, the change score is likely to be an underestimation of the actual change in locomotor disability.

Considering those in the ceiling of the distribution, changes in locomotor disability could happen in either direction to people in this group, i.e. they could improve or deteriorate. However, a change in score could only be seen for those people who deteriorated. This would lead to an impression of this group of people, on average deteriorating over time, when in fact the average change could have been an improvement, or no change at all. This issues leads to an increased risk of regression to the mean. This is a statistical phenomenon in which people initially assessed as being at the extremes of a distribution, are likely, by chance to move towards the mean of that distribution. The same principle applies to those in the floor of the distribution, in whom only an improvement can be detected. In a regression analysis, with a change score as an outcome, this could bias regression coefficients, with the size of this bias dependent on the size of the floor and ceiling effects, both in absolute terms and in relation to each other.

The problems raised by the presence of these floor and ceiling effects need to be recognised in any analysis considering changes in locomotor disability. However, analyses of the whole distribution of disability are less susceptible to these issues, as they do not directly address changes in score.

15.3.4 The use of the NorStOP datasets

Data from the NorStOP were used throughout this thesis to test the psychometric properties of items from the PF-10, form them into a measure of locomotor disability and then to test this measure and demonstrate its potential uses.

These datasets were chosen for practical reasons. The NorStOP is a large, population-based, longitudinal study for which the data were readily available, and each stage of the study not only included the PF-10, but also many other health assessment instruments. This allowed

psychometric assessment of the individual PF-10 items, as well as the new measure of locomotor disability.

A national, or even international, sample would have been ideal for the purpose of this PhD project, as it would potentially have been more generalisable. Generalisability was important in two respects in this thesis: the development of the locomotor disability measure, and the demonstration of the use of PF-10 items and the new measure in epidemiological settings.

The NorStOP sample was shown to have a higher proportion of females, and individuals from lower socioeconomic groups than the population of England as a whole, although comparisons were not always straightforward. There was also some evidence that those aged 60 to 69 years were over-represented in the sample. This was not necessarily the result of response bias. Although this is likely to have played a part, the sampling frame for the NorStOP had a slightly different structure from the English population in terms of age and socioeconomic status. Regardless of the reasons for the differences in the population structure, a higher proportion of females and individuals from lower socioeconomic groups is likely to result in a sample that reports higher levels of ill health (Department of Health 2003).

To a large extent, the development of the measure of locomotor disability should be unaffected by the overall level of ill health in the sample. This is because the Rasch model parameters, i.e. the scores of individuals and difficulties of items, can be estimated independently of the sample. However, the results of the psychometric testing of the measure, as well as the testing of the individual items, could be affected by the choice to use these data. In particular, the floor and ceiling effects are dependent on the ability of the sample. In a less able sample, for example from a clinical setting, the ceiling effect would likely be smaller. However, in the context of a population sample, the NorStOP was not so considerably different from the national population as to make the estimates of the size of these effects unhelpful.

In terms of the use of the PF-10 items and the locomotor disability measure in epidemiological settings (Chapters 8 and 14), a higher level of ill health has the potential to increase the mean level of disability in the sample, and hence to increase the proportions of people reporting limitation (Chapter 8) and shift the distribution of disability to the right (Chapter 14). There is no reason though to assume that the associations investigated within the NorStOP sample

should be inaccurate. For example, there is no reason to suspect that the difference in the mean change in locomotor disability level between those who develop pain and those who do not should differ.

Attrition from the NorStOP1 subcohort over the six-year follow-up period was substantial with only 25% of the original sample participating at all three time points (Table 4.10). Loss to follow-up was associated with age, gender, socioeconomic status and baseline health status, suggesting that the cohort suffered attrition bias, with those with in better health remaining in the study.

A healthier sample at follow-up than would be representative of the underlying population means that responsiveness is likely to have been underestimated, due to a larger ceiling effect than would have been the case in a less healthy sample. Given that responsiveness was shown to be acceptable even with this level of attrition, this should not detract from the results of the responsiveness analyses. It should be noted though that such attrition can lead to lower estimates of the level of locomotor disability in a population, and cause a left skew in the distribution.

15.3.5 Envisioning and analysing longitudinal data from the measure

The new interval-level measure of locomotor disability derived and tested in this thesis presents key advantages over the use of a binary definition of disability in a longitudinal setting.

First, locomotor disability is a dynamic process and not a fixed state (Clark et al 1998a and 1998b, Khokhar et al 2001, Wannamethee et al 2005, Gill et al 2006). Whilst a binary definition can detect changes from 'disabled' to 'not disabled' and vice versa, detection of these changes relies on changes in the level of locomotor disability over a pre-defined threshold. The measure of locomotor disability allows small changes to be detected, within the 'disabled' or 'not disabled' categories and ensures that small changes in the underlying construct that occur close to the threshold between disability and no disability are not given excess weight in analyses, as they could be with the use of a binary definition of disability.

Second and relating to cross-sectional as well as longitudinal settings, the interval-level nature of the locomotor disability measure allows the distribution of disability in the whole population to be considered, as suggested by Rose (1982 and 1985). This provides the opportunity to consider the distribution of disability now and how this might change in the future. If locomotor disability is a predictor of other diseases and disabilities, as has been suggested (Chaves et al 2000, Melzer et al 2005, Blazer et al 2006, Gill et al 2006, Peat et al 2006a), then in the same way that Rose (1985) considered that shifting the distribution of blood pressure by 10mmHg to the left would reduce attributable mortality by 30%, so could shifting the distribution of locomotor disability to the left reduce the attributable effects of this health problem.

Despite the limitations of the measure, in particular the ceiling effect, it has still been demonstrated that such a measure can add to the epidemiological investigation of locomotor disability in middle and old-age. For example, individuals who experience an onset of pain experience a concurrent increase in locomotor disability. Recovery from this pain is accompanied by a relative decrease in locomotor disability, but these individuals do not return to their original level of functioning. Assuming this increase in disability was seen as an onset of locomotor disability using a binary definition, these individuals would either appear not to recover from this disability, or would be classified as 'recovered', i.e. they no longer have disability. Neither of these

conclusions would be entirely correct, and would not give a full picture of the burden of locomotor disability in that population group.

15.4 Implications for future research

A review of their programme of work around locomotor disability has recently been published by Ebrahim et al (2008). In this review, they call for more complex models to investigate the causes of locomotor disability and the way in which these factors interact with each other. For this purpose, a flexible measure of locomotor disability will be needed. This thesis has argued for and derived an interval-level measure of locomotor disability that mimics the reality of the disability continuum and has demonstrated some aspects of its potential in investigating the epidemiology of locomotor disability.

This is not to say the binary definitions of locomotor disability should be abandoned entirely. Indeed, they are necessary for decision making, but it should be recognised that on their own, these definitions cannot provide the full picture. In an ideal situation, researchers (and clinicians) would employ an interval-level measure of locomotor disability with which to monitor individuals over time. Onto this measure a threshold of disability, for use in decision making, would be mapped. The use of Rasch analysis, with the score representing a sufficient statistic for the response pattern means that this approach would have the added advantage that a qualitative meaning could be assigned to the threshold. If necessary, different thresholds could be placed along the same continuum for different decision making purposes.

Despite the extensive testing of this new locomotor disability score in this thesis, both within the NorStOP and other populations, further work in other samples will be required in order to fully test the measure. In the immediate future, research should concentrate on the testing, including fit to the Rasch model specification, reproducibility, validity and responsiveness in other populations in which the PF-10 has been administered. The fit of the data to the Rasch model in the other datasets (Chapter 11) is promising, but on its own, cannot confirm the suitability of the measure for use in other populations. Different language and cultural adaptations may require a different scoring mechanism.

Future research would also benefit from a method by which to assign qualitative meaning to change scores. For example, a minimally important change statistic (MIC) could be developed. The point of this statistic would be purely to help with the interpretation of changes in locomotor disability score, especially in large scale epidemiological studies, where large sample sizes may regularly produce statistically significant results from small changes in disability level. Another use for such a statistic, although not considered directly in the development of the measure would be the calculation of sample size in a clinical trial. It was not possible to calculate an MIC as part of this PhD project because no suitable comparator data, including an anchor against which to judge important changes in locomotor disability, were available.

Melzer et al (2004) conducted a study to compare levels of locomotor disability across population groups, assessing differences in the reporting of locomotor disability across cultures. Due to the wide-spread use of the SF-36, the new measure of locomotor disability could be used to further this work and provide an estimate of the distribution of disability internationally. This would add considerably to an appreciation of the international differences in disability, because, as shown in Chapter 3, any comparison between population groups is problematic due to the use of different definitions of disability in different studies. Comparing populations based on distribution of disability, rather than simply prevalence estimates, would facilitate an understanding of why prevalence estimates differ and help to quantify these differences further.

This thesis has argued for the use of an interval-level measure of locomotor disability in using selected items from the PF-10. It remains, however, to demonstrate directly its performance in relation to both the original PF-10 and a summated ratings approach to the five locomotor disability-specific items selected for the new measure. The performance of the new measure in relation to these two potential alternatives should be tested psychometrically, including the repeatability, validity and feasibility, and also in terms of the conclusions of studies using these measures, by for example repeating the analyses in Chapter 14. Differences in the psychometric performance of the scales might provide information as to the most appropriate method of scoring. Different conclusions from studies using the summated ratings score from the five locomotor disability-specific items and the new locomotor disability measure would suggest that the interval-level score gained from the Rasch methodlogy presents not just a theoretically more correct

method of scale derivation, but also a score that is different in practical terms from more traditionally constructed scores. Different conclusions from the new measure and the original PF-10 could suggest a similar methodological point, but may also provide empirical evidence in favour of the use of only five of the PF-10 items to create the new measure.

A drawback of the use of the new measure of locomotor disability, in particular for clinical practice, is the need for Rasch analysis software, such as RUMM2020 (Andrich et al 2003) or Winsteps (Winsteps.com 2009), in order to create the score from the super-items. The creation of the score requires both the software and the expertise to use it: neither is likely to be widely available. Furthermore, the method employed throughout this thesis, of using a Rasch analysis program to create the score afresh in each sample, results in a slightly different score for the same response to the PF-10 items on each occasion. For example, in the NorStOP pilot dataset in which the score was developed, those in the ceiling of the distribution (no disability) scored -4.213 logits, whilst in the WHS people in the ceiling were allocated a score of -3.378 logits. This situation has been likened to the weights and measures used in eighteenth century Parisian markets. These measures were consistent within, but not across markets. Trade between markets was not possible until the measures were standardised across these markets (Trevor Bond personal communication, 15 December 2009). The same is true of these Rasch-based scoring mechanisms, where scores need to be anchored over studies, as scores were anchored over time in this thesis. Further into the future, research in the field of measurement should concentrate on the transfer of the score between different samples because this inconsistency in scoring creates confusion and is likely to discourage the use of this measure of locomotor disability and other scores created using the Rasch model. This problem can, on the face of it, be overcome through the use of a conversion table, rather than applying the Rasch model in each sample, or by transforming the score onto another scale, say 0 to 100. However, both of these approaches have their problems. As discussed in Chapter 9, a conversion table is straightforward, but is not considered to be appropriate (Wright 2003): it is likely to result in the loss of interval-level measurement, the major advantage of the Rasch methodology. Transformation of the score to another scale merely hides the problem: it does not solve it, although it might make the Rasch score more appealing to other researchers. Bond and Fox (2001) suggest that it is possible to pre-specify the values of item

locations in the Rasch model estimation and therefore to maintain a fixed ruler across samples. However, this means assuming that item locations would be approximately the same across populations, and this is not necessarily the case if the population itself is fundamentally different.

The advantages of an interval-level measure are not confined to locomotor disability and similar methods should be employed in other fields where the construct of interest is continuous. This should particularly be the case if a method is found to ensure a score that is transferable between samples, as discussed above. The drawback with this idea is that previously constructed tools, for example the Western Ontario and McMaster Osteoarthritis Index (Bellamy et al 1988) for knee and hip pain, or the Roland and Morris Disability Questionnaire for back pain (Roland & Morris 1983), will often not fit the Rasch model specification and so no interval-level measure can be created (Davis et al 2003, Garratt et al 2003). It would however be sensible for those developing new questionnaires to employ Rasch model methodology in order to select well-functioning items that form a single dimension and will give an interval-level summary measure. This has been done previously in some areas of research, for example foot disability (Redmond et al 2006) and this should continue and be developed further, both in disability-related fields and more widely.

Although the locomotor disability measure has its limitations and further work will be required before it can be widely used, it allows locomotor disability to be assessed as a spectrum, which is more natural both in terms of its cross-sectional nature and the ability to quantify changes in the level of disability. This should add to the research called for by Ebrahim et al (2008) to consider methodologically more rigorous approaches to the research of locomotor disability.

References

Adamson J, Hunt K, Ebrahim S. Socioeconomic position, occupational exposures, and gender: the relation with locomotor disability in early old age. Journal of Epidemiology and Community Health 2003; 57: 453-455

Adamson J, Lawlor DA, Ebrahim S. Chronic diseases, locomotor activity limitation and social participation in older women: cross sectional survey of British Women's Heart and Health Study. Age and Ageing 2004; 33: 293-298

Ahacic K, Parker MG, Thorslund M. Mobility limitations in the Swedish population from 1968 to 1992: Age, gender and social class differences. Aging 2000; 12: 190-198

Ahacic K, Parker MG, Thorslund M. Mobility limitations 1974-1991: period changes explaining improvement in the population. Social Science & Medicine 2003; 57: 2411-2422

Äijänseppä S, Notkola IL, Tijhuis M, van Staveren W, Kromhout D, Nissinen A. Physical Functioning in elderly Europeans: 10 year changes in the north and south: the HALE project. Journal of Epidemiology and Community Health 2005; 59: 413-419

Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1991

Ananth CV, Kleinbaum DG Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology 1997; 26: 1323-1333

Anderson JA. Regression and ordered categorical variables. Journal of the Royal Statistical Society Series B-Methodological 1984; 46: 1-30

Andresen EM, Bowley RN, Rothenberg BM, Panzer R, Katz P. Test-retest performance of a mailed version of the Medical Outcomes Study 36-item short-form health survey among older adults. Medical Care 1996; 34: 1165-1170

Andresen EM, Rothenberg BM, Panzer R, Katz P, McDermott MP. Selecting a generic measure of health-related quality of life for use among older adults. A comparison of candidate instruments. Evaluation and the Health Professions 1998; 21: 244-264

Andrews GR, Esterman AJ, Braunack-Mayer AJ, Rungie CM. Aging in the Western Pacific. Manila: WHO; 1986

Andrich D. A rating formulation for ordered response categories. Psychometrika 1978; 43: 561-573

Andrich D. An elaboration of Guttman scaling with Rasch models for measurement. In Sociological Methodology. Edited by Brandon-Tuma N. San Fransisco: Jossey-Bass; 1985: pg 33-80

Andrich D, Lyne A, Sheridan B, Luo G. RUMM2020. Perth: RUMM Laboratory; 2003

Andrich D. Controversy and the Rasch Model. A characteristic of incompatible paradigms? Medical Care 2004; 42: I7-I16

Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. American Journal of Epidemiology 1989; 129: 191-204

Avlund K, Luck M, Tinsley R. Cultural differences in functional ability among elderly people in Birmingham, England, and Glostrup, Denmark. Journal of Cross-cultural Gerontology 1996; 11: 1-16

Avlund K, Damsgaard MT, Sakari-Rantala R, Laukkanen P, Schroll M. Tiredness in daily activities among nondisabled old people as a determinant of onset of disability. Journal of Clinical Epidemiology 2002; 55: 965-973

Avlund K, Vass M, Hendriksen C. Onset of mobility disability among community-dwelling old mean and women. The role of tiredness in daily activities. Age and Ageing 2003; 32: 579-584

Bajekal M, Primatesta P, Prior G (Editors). Health Survey for England 2001: Disability. London: The Stationary Office; 2003

Bannerman E, Miller MD, Daniels LA, Cobiac L, Giles LC, Whitehead C, Andrews GR, Crotty M. Anthropometric indices predict physical function and mobility in older Australians: the Australian Longitudinal Study of Ageing. Public Health Nutrition 2002; 5: 655-662

Barnett AG, van der Pols J, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology 2005; 34: 215-220

Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. Journal of Rheumatology 1988; 15: 1833-40

Bejar II. Achievement Testing: Recent Advances. Beverly Hills: Sage Publications; 1983

Bender R, Grouven U. Ordinal logistic regression in medical research. Journal of the Royal College of Physicians of London 1997; 31: 546-551

Bender R, Grouven U. Using binary logistic regression models for ordinal data with non-proportional odds. Journal of Clinical Epidemiology 1998; 51: 809-816

Berglund A, Bodin L, Jensen I, Wiklund A, Alfredsson L. The influence of prognostic factors on neck pain intensity, disability, anxiety and depression over a 2-year period in subjects with acute whiplash injury. Pain 2006; 125: 244-256

Bergner M, Bobbit RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Medical Care 1981; 19: 787-805

Berkman LF, Syme SL. Social networks, Host resistance and mortality: a nine-year follow-up study of Alameda County residents. American Journal of Epidemiology 1979; 109: 186-204

Bhuiyan AR, Gustat J, Srinivasan SR, Berenson GS. Differences in body shape representations among young adults from a biracial Black-White, semirural community: the Bogalusa Heart Study. American Journal of Epidemiology 2003; 158: 792-797

Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties and clinical application. Physical Therapy 1999; 79: 371-383

Birrell F, Croft P, Cooper C, Hosie G, Macfarlane GJ, Silman A. Radiographic change is common in new presenters in primary care with hip pain. Rheumatology 2000; 39: 772-775

Blazer DG, Hybels CF, Fillenbaum GG. Metabolic Syndrome predicts mobility decline in a community-based sample of older adults. Journal of the American Geriatrics Society 2006; 54: 502-506

Boardman HF, Thomas E, Millson DS, MacGregor EA, Laughey WF, Croft PR. North Staffordshire Headache Survey: development, reliability and validity of a questionnaire for use in a general population survey. Cephalgia 2003; 23: 325-331

Bohannon RW, Brennan PJ, Pescatello LS, Marschke L, Hasson S, Murphy M. Using self-report and speed to screen for gait limitations. Physical and Occupational Therapy in Geriatrics 2004; 23: 1-8

Bombardier C, Tugwell P. Methodological considerations in functional assessment. Journal of Rheumatology 1987; 14 (Supp15): 6-10

Bond TG, Fox CM. Applying the Rasch model. Mahwah: Lawrence Erlbaum Associates; 2001

Bond TG. Measurement scales are for measuring patients: where Rasch measurement should be going in rehabilitation research. European Journal of Physical and Rehabilitation Medicine 2008; 44: 359-363

Bot SDM, Terwee CB, van der Windt DAWM, Bouter LM, Dekker J, de Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Annals of the Rheumatic Diseases 2004; 63: 335-341

Bowling A, Bond M, Jenkinson C, Lamping DL. Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. Journal of Public Health 1999; 21: 255-270

Bowling A. research methods in health. Investigating health and health services. 2nd edition. Buckingham: Open University Press; 2002

Brach JS, Studenski SA, Perera S, VanSwearingen JM, Newman AB. Gait variability and the risk of incident mobility disability in community-dwelling older adults. The Journals of Gerontology: Series A, Biological Sciences and Medical Sciences 2007; 62: 983-988

Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, Ware JE. Better assessment of physical function: item improvement is neglected but essential. Arthritis Research & Therapy 2009; 11: R191

Byrne BM. Structural Equation Modeling with AMOS. Basic concepts, applications and programming. Mahwah: Lawernce Erlbaum Associates; 2001

Campbell MJ. Statistics at Square Two: Understanding Modern Statistical Applications in Medicine. London: BMJ Books; 2001

Cardiff Software Inc. Teleform Standard User guide Version 7.0. San Marcos: Cardiff Software; 1998

Chang M, Cohen-Mansfield J, Ferrucci L, Leveille S, Valpato S, de Rekeneire N, Guralnik JM. Incident loss of ability to walk 400 meters in a functionally limited older population. Journal of the American Geriatrics Society 2004; 52: 2094-2098

Chaves PHM, Garrett ES, Fried LP. Predicting the risk of mobility difficulty in older women with screen nomograms. Archives of Internal Medicine 2000; 160: 2525-2533

Cieza A, Brockow T, Ewert T, Amman E, Kollertis B, Chatterji S, Üstün B, Stucki G. Linking health-status measurements to the international classification of functioning, disability and health. Journal of Rehabilitation Medicine 2002; 34: 205-210

Clark DO, Stump TE, Wolinsky FD. Predictors of onset and recovery from mobility difficulties among adults aged 51-61 years. American Journal of Epidemiology 1998a; 148: 63-71

Clark DO, Stump TE, Hui SL, Wolinsky FD. Predictors of mobility and basic ADL difficulty among adults aged 70 years and over. Journal of Aging and Health 1998b; 10: 422-440

Clogg CC. Latent class models. In Handbook of statistical modeling for the social and behavioural sciences. Edited by Arminger G, Clogg CC & Sobel ME. New York: Plenum Press; 1995: 331-359

Cole SR. A reanalysis of data from regression models for ordinal responses'. International Journal of Epidemiology 1999; 28: 805

Coppin AK, Ferrucci L, Lauretani F, Phillips C, Chang M, Bandinelli S, Guralnik JM. Low socioeconomic status and disability in old age: evidence from the InChianti study for the mediating role of physiological impairments. The Journal of Gerontology. Series A, Biological Sciences and Medical Sciences 2006; 61: 86-91

Croft P, Jordan K, Jinks C. "Pain elsewhere" and the impact of knee pain in older people. Arthritis and Rheumatism 2005; 52: 2350-2354

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297-334

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. Journal of Clinical Epidemiology 2003; 56: 395-407

Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, Williams JI. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. Journal of Clinical Epidemiology 2003; 56: 1076-1083

de Vet HCW, Terwee CB, Bouter LM. Current challenges in clinimetrics. Journal of Clinical Epidemiology 2003; 56: 1137-1141

de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability parameters. Journal of Clinical Epidemiology 2006; 59: 1033-1039

de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, Boers, M, Bouter LM. Minimally important change determined by a visual method integrating an anchor-based and distribution-based approach. Quality of Life Research 2007; 16: 131-142

Department of Health. Tackling Health Inequalities: A Programme for Action. London: Department of Health; 2003

Department of Health. Health Survey for England 2005. London: Department of Health; 2007

Department of Health And Human Services. Centers for Disease Control and Prevention; 2007 Available at: http://www.cdc.gov/nccdphp/dnpa/bmi/adult_BMI/about_adult_BMI.htm last accessed 13 September 2010

Dubuc N, Haley SM, Kooyoomjian JT, Jette AM. Assessing disability in older adults: the effects if asking questions with and without health attribution. Journal of Rehabilitation Medicine 2004; 36: 226-231

Dunlop DD, Hughes SL, Manheim LM. Disability in activities of dialing living: patterns of change and a hierarchy of disability. American Journal of Public Health 1997; 87: 378-383

Dunn KM, Jordan K, Croft PR. Characterizing the course of low back pain: a latent class analysis. American Journal of Epidemiology 2006; 163: 154-761

Dutta C, Hadley EC. The significance of sarcopenia in old age. The Journals of Gerontology: Series A, Biological Sciences and Medical Sciences 1995; 50: S1-S4

Ebrahim S, Wannamethee SG, Whincup P, Walker M, Shaper AG. Locomotor disability in a cohort of British men: the impact of lifestyle and disease. International Journal of Epidemiology 2000; 29: 478-486

Ebrahim S, Adamson J, Ayis S, Beswick A, Gooberman-Hill R. Locomotor disability: meaning, causes and effects of interventions. Health Services Research & Policy 2008; 13 Supp 3: 38-46

Epachnikov VA. Nonparametric estimation of a multidimensional probability density. Theory of Probability and its Applications 1969; 14: 156–161

Evandrou M. Health and social care. In Focus on older people. Edited by Soule A, Babb P, Evandrou M, Balchin S, Zealey L. London: HMSO; 2005a: 51-65

Evandrou M. Health and well being. In Focus on older people. Edited by Soule A, Babb P, Evandrou M, Balchin S, Zealey L. London: HMSO; 2005b: 39-50

Eyres S, Carey A, Gilworth G, Neumann V, Tennant A. Construct validity and reliability of the Rivermead Post-Consussion Symptoms Questionnaire. Clinical Rehabilitation 2005; 19: 878-887

Feinstein AR, Josephy B R, Wells C K. Scientific and clinical problems in indexes of functional disability. Annals of Internal Medicine 1986; 105: 413-420

Fischer GH. Derivations of the Rasch model. In Rasch models: foundations, recent developments, and applications. Edited by Fischer GH, Molenaar IW. New York: Springer-Verlag; 1995; pg 15-38

Flegal KM, Troiana RP. Changes in the distribution of body mass index of adults and children in the US population. International Journal of Obesity 2000; 24: 807-818

Fletcher RW, Fletcher SW. Clinical Epidemiology: The Essentials. 4th edition. Baltimore: Lippincott Williams & Wilkins; 2005

Foreman MD, Kleinpell R. Assessing the quality of life of elderly persons. Seminars in Oncology Nursing 1990; 6: 292-297

Forrest M, Andersen B. Ordinal scale and statistics in medical research. British Medical Journal 1986; 282: 537-8

Fried LP, Bandeen-Roche K, Chaves PHM, Johnson BA. Preclinical mobility disability predicts incident mobility disability in older women. The Journals of Gerontology: Series A, Biological Sciences and Medical Sciences 2000; 55: M43-M52

Fried LP, Guralnik JM. Disability in older adults: evidence regarding significance, etiology, and risk. Journal of the American Geriatrics Society 1997; 45: 92-100

Fries JR. Aging, natural death, and the compression of morbidity. New England Journal of Medicine 1980; 303: 130-135

Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. The Journal of Rheumatology 1982; 9: 789-793

Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progess in assessing physical functionin artritis: PROMISshort forms and computerized adaptive testing. Journal of Rheumatology 2009; 36: 2061-2066

Gandek B, Ware JE, Aaronson NK, Alonso J, Apolone G, Bjorner J, Brazier J, Bullinger M, Fukuhara S, Kaasa S, Leplège A, Sullivan M. Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the IQOLA Project. Journal of Clinical Epidemiology 1998; 51: 1149-1158

Gardener EA, Huppert FA, Guralnik JM, Melzer D. Middle-aged and mobility-limited. Prevalence of disability and symptoms attribution in a national survey. Journal of General Internal Medicine 2006; 21: 1091-1096

Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF 36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? British Medical Journal 1993; 306: 1440-1444

Garratt A, Schmidt L, Macintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. British Medical Journal 2002; 324: 1417-1420

Garratt AM, United Kingdom Back Pain Exercise and Manipulation Trial. Rasch analysis of the Roland Disability Questionnaire. Spine 2003; 28: 79-84

Gershon RC. Computer adaptive testing. In Introduction to Rasch Measurement. Edited by Smith EV, Smith RM. Maple Grove: JAM Press; 2004: pg 601-629

Gill TM, Allore HG, Hardy SE, Guo Z. The dynamic nature of mobility disability in older persons. The Journal of the American Geriatrics Society 2006; 54: 248-254

Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. When is baseline adjustment useful in analysis of change? An example with education and cognitive change. American Journal of Epidemiology 2005; 162: 267-287

Greenland S. Alternative models for ordinal logistic regression. Statistics in Medicine 1994; 13: 1665-1677

Gregory PC, Fried LP. Why do older adults decide they are having difficulty with a task? American Journal of Physical Medicine & Rehabilitation 2003; 82: 9-16

Grundy E, Holt G. The socioeconomic status of older adults: How should we measure it in studies of health inequalities? Journal of Epidemiology and Community Health 2001; 55: 895-904

Guralnik JM, LaCroix AZ, Abbott RD, Berkman LF, Scatterfield S, Evans DA, Wallace RB. Maintaining mobility in late life: I. Demographic characteristics and chronic conditions. American Journal of Epidemiology 1993; 137: 845-857

Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, Scherr PA, Wallace RB. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. Journal of Gerontology 1994; 49: M85-94

Guralnik JM, Fried LP, Salive ME. Disability as a public health outcome in the aging population. Annual Review of Public Health 1996; 17: 25-46

Guralnik JM, Ferrucci L, Pieper CF, Leveille SG, Markides KS, Ostir GV, Studenski S, Berkman LF, Wallace RB. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. The Journals of Gerontology: Series A, Biological Sciences and Medical Sciences 2000; 55: M221-M231

Guralnik JM, Ferrucci L, Balfour JL, Volpato S, Di Iorio A. Progressive versus catastrophic loss of the ability to walk: implications for the prevention of mobility loss. Journal of the American Geriatrics Society 2001; 49: 1463-1470

Guralnik JM, Leveille S, Volpato S, Marx MS, Cohen-Mansfield J. Targeting high risk older persons into exercise programs for disability prevention. Journal of Aging and Physical Activity 2003; 11: 219-228

Gureje O, Simon GE, Von Korff M. A cross-national study of the course of persistent pain in primary care. Pain 2001; 92: 195-200

Haley SM, McHorney CA, Ware JE. Evaluation of the MOS SF-36 physical functioning scale PF-10: I Unidimensionality and reproducibility of the Rasch item scale. Journal of Clinical Epidemiology 1994; 47: 671-684

Harwood RH. Locomotor disability. In Epidemiology in Old Age. Edited by Ebrahim S, Kalache A. London: BMJ Publishing Group; 1996: 378-388

Hayes V, Morris J, Wolfe C, Morgan M. The SF-36 health survey questionnaire: is it suitable for use with older adults? Age and Ageing 1995; 24: 120-125

Heikkinen E, Waters WE, Brzezinski ZJ. The elderly in eleven countries. A sociomedical survey. Copenhagen: WHO; 1983

Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 1965; 58: 295-300

Hirani V, Malbut K. Disability among older people. In Health Survey for England 2000. Edited by Prior G, Primatesta P. London: The Stationary Office; 2002

Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technology Assessment 2009; 13: 1-177

Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons; 2000

Wennie Huang WN, Perera S, VanSwearingen J, Studenski S. Performance measures predict onset of activity of daily living difficulty in community-dwelling older adults. Journal of the American Geriatrics Society 2010; 58: 844-852

Iberg KM, Salomom JA, Tandon A, Murray CJL. Cross-population comparability of self-reported and physician-assessed mobility levels: Evidence from the Third National Health and Nutrition Examination Survey. Global Programme on Evidence for Health Policy Discussion Paper No 14. Geneva: WHO; 2001

Iezzoni LI, McCarthy EP, Davis RB, Siebens H. 2001 Mobility difficulties are not only a problem of old age. Journal of General Internal Medicine 2001; 16: 235-24

Jagger C, Matthews R, Matthews F, Robinson T, Robine JM, Brayne C, Medical Research Council Cognitive Function and Ageing Study Investigators. The burden of diseases on disability-free life expectancy in later life. The Journals of Gerontology: Medical Sciences 2007a; 62A: M408-M414

Jagger C, Matthews RJ, Matthews FE, Spiers NA, Nickson J, Paykel ES, Huppert FA, Brayne C, MRC-CFAS team. Cohort differences in disease and disability in the young-old: findings from the MRC Cognitive Function and Ageing Study (MRC-CFAS). BMC Public Health 2007b; 7: 156

Jenkinson C, Fitzpatrick R, Garrett A, Peto V, Stewart-Brown S. Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). Journal of Neurology, Neurosurgery and Psychiatry 2001; 71: 220-224

Jinks C, Jordan K, Ong BN, Croft P. A brief screening tool for knee pain in primary care (KNEST). 1. Validity and reliability. Rheumatology 2004; 40: 528-536

Jinks C, Jordan K, Ong BN, Croft P. A brief screening tool for knee pain in primary care (KNEST).2. Results from a survey in the general population aged 50 years and over. Rheumatology (Oxford) 2004; 43: 55-51

Karabatsos G. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. In Introduction to Rasch Measurement. Edited by Smith EV, Smith RM. Maple Grove: JAM Press; 2004: 630-634

Keenan AM, Redmond AC, Horton M, Conaghan PG, Tennant A. The Foot Posture Index: Rasch analysis of a novel, foot-specific outcome measure. Archives of Physical Medicine and Rehabilitation 2007; 88: 88-93

Khokhar SR, Stern Y, Bell K, Anderson K, Noe E, Mayeux R, Albert S. Persistent mobility deficit in the absence of deficits in activities of daily living: a risk factor for mortality. Journal of the American Geriatrics Society 2001; 49: 1539-1543

Koster A, Penninx BW, Newman AB, Visser M, van Gool CH, Harris TB, van Eijk, Kempen, GI, Brach JS, Simonsick EM, Houston DK, Tylavsky FA, Rubin SM, Kritchevsky SB. Lifestyle factors and incident mobility limitation in obese and non-obese older adults. Obesity (Sliver Spring) 2007; 15: 3122-3132

Kriegsman DMW, Deeg DJH, van Eijk JTM; Penninx BWJ, Boeke A. Do disease specific characteristics add to the explanation of mobility limitations in patients with different chronic diseases? A study in the Netherlands. The Journal of Epidemiology and Community Health 1997; 51: 676-685

Lacey RJ, Lewis M, Jordan K, Jinks C, Sim J. Inter-rater reliability assessment of the scoring of the body pain manikin. Spine 2005; 30: E445-448

LaCroix AZ, Guralnik JM, Berkman LF, Wallace RB, Scatterfield S. Maintaining mobility in late life. II. Smoking, alcohol consumption, physical activity, and body mass index. American Journal of Epidemiology 1993; 137: 858-869

Lall R, Campbell MJ, Walters SJ, Morgan K. A review of ordinal regression models applied on health-related quality of life assessments. Statistical Methods in Medical Research 2002; 11: 49-67

Lan T-Y, Melzer D, Tom BD, Guralnik JM. Performance tests and disability: developing an objective index of mobility-related limitation in older populations. The Journals of Gerontology: Medical Sciences 2002; 57A: M294-M301

Launer LJ, Harris T, Rumpel C, Madans J. Body mass index, weight change, and risk of mobility disability in middle-aged and older women. The epidemiologic follow-up study of NHANES I. Journal of the American Medical Association 1994; 271: 1093-1098

Leveille SG, Penninx BWJH Melzer D, Izmirlian G, Guralnik JM. Sex differences in the prevalence of mobility disability in old age: the dynamics of incidence, recovery, and mortality. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences 2000; 55: S41-S50

Leveille SG, Fried L, Guralnik JM. Disabling symptoms: what do older women report? Journal of General Internal Medicine 2002; 17: 766-773

Leveille SG, Bean J, Ngo L, McMullen W, Guralnik JM. The pathway from musculoskeletal pain to mobility difficulty in older disabled women. Pain 2007; 128: 69-77

Li Y. Using the open-source statistical language R to analyze the dichotomous Rasch model. Behavioral Research Methods 2006; 38: 532-541

Linacre JM. Sample size and item calibration stability. Rasch Measurement Transactions 1994; 7: 328

Linacre JM. Optimizing rating scale category effectiveness. In Introduction to Rasch Measurement. Edited by Smith EV, Smith RM. Maple Grove: JAM Press; 2004: 258-278

Linacre JM. WINSTEPS Rasch measurement computer program: version 3.64.1. Chicago: Winsteps.com; 2007

Lohr KN, Aaronson KN, Alonso J, Burnam MA, Patrick DL, Perrin EB, Roberts JS. Evaluating quality of life and health status instruments: development of scientific review criteria. Clinical Therapeutics 1996; 18: 979-992

Long SJ, Freese J. Regression Models for Categorical Dependent Variables Using Stata. College Station: Stata Press; 2006

Lunt M. Predicting an Ordinal Outcome: Options and Assumptions. Presented at 7th UK Stata Users Group Meeting; 2001a Available at: http://www.stata.com/meeting/7uk/lunt1.pdf last accessed 13 September 2010

Lunt M. Stereotype Ordinal Regression. Stata Technical Bulletin 2001b; 61:12-18

Lunt M. Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. Statistics in Medicine 2005; 24: 1357-1369

Mallinson S. The Short-Form 36 and older people: some problems encountered when using postal administration. Journal of Epidemiology and Community Health 1998; 52: 224-328

Mallinson S. Listening to respondents: a qualitative assessment of the Short-Form 36 health status questionnaire. Social Science & Medicine 2002; 54: 11-21

Manor O, Matthews S, Power C. Dichotomous or categorical response? Analysing self-rated health and lifetime social class. International Journal of Epidemiology 2000; 29: 149-157

Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. Journal of Applied Measurement 2008; 9: 200-215

Marshall M, van der Windt D, Nicholls E, Myers H, Hay E, Dziedzic K. Radiographic hand osteoarthritis: patterns and associations with hand pain and function in a community-dwelling sample. Osteoarthritis and Cartilage 2009; 17: 1440-1447

Martin J, Meltzer H, Elliot D. OPCS surveys of disability in Great Britain. The Prevalence of disability among adults. London: HMSO; 1988

Masters G. A rasch model for partial credit scoring. Psychometrika 1982; 42: 149-174

Matthews RJ, Smith LK, Hancock RM, Jagger C, Spiers NA. Socioeconomic factors associated with the onset of disability in older age: a longitudinal study of people aged 75 years and over. Social Science & Medicine 2005; 61: 1567-1575

McCullagh P, Nelder J. Generalized Linear Models. Cambridge: Chapman and Hall; 1983

McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Medical Care 1994; 32: 40-66

McHorney CA. Measuring and monitoring general health status in elderly persons: practical and methodological issues in using the SF-36 health survey. The Gerontologist 1996; 5: 571-583

McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. Journal of Clinical Epidemiology 1997; 50: 451-461

Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2 The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis and Rheumatism 1992; 35: 1-10

Melzer S, Lan TY, Guralnik JM. The predictive validity for mortality of the index of mobility-related limitation – results from the EPESE study. Age and Ageing 2003; 32: 619-625

Melzer D, Parahyba MI. Socio-demographic correlates of mobility disability in older Brazilians: results of the first national survey. Age and Ageing 2004; 33: 253-259

Melzer D, Lan T-Y, Tom BDM, Deeg DJH, Guralnik JM. Variation in thresholds for reporting mobility disability between national population subgroups and studies. The Journals of Gerontology: Medical Sciences 2004; 59A: M1295-M1303

Melzer D, Gardener E, Guralnik JM. Mobility disability in the middle-aged: cross-sectional associations in the English Longitudinal Study of Ageing. Age and Ageing 2005; 34: 594-602

Menard S. Longitudinal Research. Thousand Oaks: Sage Publications; 2002

Mendes de Leon CF, Hansberry MR, Bienias JL, Morris MC, Evans DA. Relative weight and mobility: a longitudinal study in a biracial population of older Americans. Annals of Epidemiology 2006; 16: 770-776

Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Archives of Physical Medicine and Rehabilitation 1989; 70: 308-312

Merrill SS, Seeman TE, Kasl SV, Berkman LF. Gender differences in the comparison of self-reported disability and performance measures. The Journals of Gerontology: Medical Sciences 1997; 52A: M19-M26

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN checklist for assessing the methodological properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010; 19: 539-549

Michael YL, Berkman LF, Colditz GA, Kawachi I. Living arrangements, social integration, and change in functional health status. American Journal of Epidemiology 2001; 153: 123-131

Mor V. The compression of morbidity hypothesis: a review of research and prospects for the future. Journal of the American Geriatrics Society 2005; 53: S308-S309

Moss-Morris R, Weinman J, Petrie K, Horne R, Cameron L, Buick D. The revised illness perception questionnaire (IPQ-R). Psychology and Health 2002; 17: 1-16

Muthén, L.K. and Muthén, B.O. Mplus User's Guide. Sixth Edition. Los Angeles: Muthén & Muthén; 1998-2010

Muthén LK, Muthén BO. Mplus Version 5.21 Demo. Los Angeles: Muthén & Muthén; 2010. Accessed at http://www.statmodel.com/demo.shtml. Last accessed 13 September 2010

Myers H, Nicholls E, Handy J, Peat G, Thomas E, Duncan R, Wood L, Marshall M, Tyson C, Hay E, Dziedzic K. The Clinical Assessment Study of the Hand (CAS-HA): a prospective study of musculoskeletal hand problems in the general population. BMC Musculoskeletal Disorders 2007; 8: 85

Nagi SZ. Some conceptual issues in disability and rehabilitation. In Sociology and Rehabilitation. Edited by MB Sussman. Washington: American Sociological Association; 1965

National Centre for Social Research. Welsh Health Survey. UK Data Archive [distributor]; Colchester, Essex: 1998-2008

National Institute for Health and Clinical Excellence. Hypertension: management of hypertension in adults in primary care. NICE Clinical Guideline 34; June 2006

Nordstrom CK, Diez Roux AV, Schluz R, Haan MN, Jackson SA, Balfour JL. Socioeconomic position and incident mobility impairment in the Cardiovascular Health Study. BMC Geriatrics 2007; 7: 11

O'Connell AA. Logistic regression models for ordinal response variables. Thousand Oaks: Sage Publications; 2006

Odding E, Valkenburg HA, Algra D, Vandenouweland FA, Grobbee DE, Hofman A. Association of locomotor complaints and disability in the Rotterdam study. Annals of Rheumatic Diseases 1995; 54: 721-725

Office for National Statistics. Standard occupational classification 2000, Vol. 2. The coding index. London: The Stationery Office; 2000

Office for National Statistics. The National Statistics Socio-economic classification user manual. Version 1, 1. London: The Stationery Office; 2002

Office for National Statistics. People aged 65 and over: Results of a study carried out by the Department of Health as part of the 2001 General Household Survey. London: The Stationary Office; 2003

Office for National Statistics. Health Expectancy. Living longer, more years in poor health. Available at http://www.statistics.gov.uk/cci/nugget.asp?id=934. First published 1 October 2008. Last accessed 13 September 2010

Onder G, Penninx BW, Ferrucci L, Fried LP, Guralnik JM, Pahor M. Measures of physical performance and risk for progressive and catastrophic disability: results from the Women's Health

and Aging Study. The Journals of Gerontology: Series A, Biological Sciences and Medical Sciences 2005; 60: 74-79

Pae EK. Measurements must be interval, not ordinal. The Angle Orthodontist 1999; 69: 397

Pallent JF, Miller RL, Tennant A. Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. BMC Psychiatry 2006; 6: 28

Papageorgiou A, Croft PR, Ferry S, Jayson MI, Silman AJ. Estimating the prevalence of low back pain in the general population. Evidence from the South Manchester Back Pain Survey. Spine 1995; 20: 1889-1893

Parker SG, Peet SM, Jagger C, Farhan M, Castleden CM. Measuring health status in older patients. The SF-36 in practice. Age and Ageing 1998; 27: 13-18

Peat G, Thomas E, Handy J, Wood L, Dziedzic K, Myers H, Wilkie R, Duncan R, Hay E, Hill J, Croft P. The Knee Clinical Assessment Study – CAS(K). A prospective study of knee pain and knee osteoarthritis in the general population. BMC Musculoskeletal Disorders 2004; 5: 4

Peat G, Thomas E, Wilkie R, Croft P. Multiple joint pain and lower extremity disability in middle and old age. Disability and Rehabilitation 2006a; 28: 1543-1549

Peat G, Thomas E, Handy J, Wood L, Dziedzic K, Myers H, Wilkie R, Duncan R, Hay E, Hill J, Lacey R, Croft P. The Knee Clinical Assessment Study – CAS(K). A prospective study of knee pain and knee osteoarthritis in the general population: baseline recruitment and retention at 18 months. BMC Musculoskeletal Disorders 2006b, 7: 30

Pfeiffer E. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. Journal of the American Geriatrics Society 1975; 23: 433-441

Pollard B, Johnston M, Dixon D. Theoretical framework and methodological development of common subjective health outcome measures in osteoarthritis: a critical review. Health and Quality of Life Outcomes 2007; 5: 14

Porter EJ. Scales and tales: older women's difficulty with daily tasks. The Journals of Gerontology: Series B Social Sciences 2007; 62B: S153-S159

Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. Journal of Clinical Epidemiology 1998; 51: 1203-1214

Rasch G. Probabilistic models for some intelligence and attainment tests. 1960 [Reprinted by University of Chicago Press, 1980]

Redmond AC, Crosbie J, Ouvrier RA. Development and validation of a novel rating system for scoring standing foot posture: The Foot Posture Index. Clinical Biomechanics 2006, 21: 89-98

Reichenheim M. Confidence intervals for the Kappa statistic. The Stata Journal 2004; 4: 421-428

Roland M, Morris M. A study of the natural history of back pain: 1. Development of a reliable and sensitive measure of disability in low back pain. Spine 1983; 8: 141-144

Rose G. Incubation period of coronary heart disease. British Medical Journal 1982; 284: 1600-1601

Rose G. Sick individuals and sick populations. International Journal of Epidemiology 1985; 14: 32-38

Rose G. The strategy of preventive medicine. Oxford: Oxford University Press; 1992

Rosow I, Breslau N. A Guttman health scale for the aged. Journal of Gerontology 1966; 21: 566-559

Rowlands S, Moser K. Consultation rates from the General Practice Research Database. British Journal of General Practice 2002; 52: 658-660

Sainio P, Koskinen S, Heliövaara M, Martelin T, Härkänen T, Hurri H, Miilunpalo S, Aromaa A. Self-reported and test-based mobility limitations in a representative sample of Finns aged 30+. Scandinavian Journal of Public Health 2006; 34: 378-386

Sainio P, Martelin T, Koskinen S, Heliövaara M. Educational differences in mobility: the contribution of physical workload, obesity, smoking and chronic conditions. Journal of Epidemiology and Community Health 2007; 61: 401-408

Sakari-Rantala R, Avlund K, Frändin K, Era P. The incidence of mobility restrictions among elderly people in two Nordic localities. A five-year follow-up. Aging Clinical and Experimental Research 2002; 14: 47-55

Salgado-Ugarte IH, Shimizu M, Taniuchi T. Snp6: Exploring the shape of univariate data using kernal density estimators. Stata Technical Bulletin 1993; 16: 8-19

Salive ME, Guralnik J, Gylnn RJ, Christen W, Wallace RB, Ostfeld AM. Association of visual impairment with mobility and physical function. Journal of the American Geriatrics Society 1994; 42: 287-292

Sayers SP, Brach JS, Newman AB, Heeren TC, Guralnik JM, Fielding RA. Use of self-report to predict ability to walk 400 meters in mobility-limited older adults. Journal of the American Geriatrics Society 2004; 52: 2099-2103

Seeman TE. Social ties and health: the benefits of social integration. Annals of Epidemiology 1996; 6: 442-445

Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. Journal of Clinical Epidemiology 1997; 50: 45-55

Shumway-Cook A, Ciol MA, Yorkston KM, Hoffman JM, Chan L. Mobility limitations in the Medicare population: prevalence and sociodemographic and clinical correlates. Journal of the American Geriatrics Society 2005; 53: 1217-1221

Sim J, Wright CC. Kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 2005; 85: 257-268

Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principle components analysis of residuals. Journal of Applied Measurement 2002; 3: 205-231

Smith AB, Rush R, Fallowfield AJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology 2008; 8: 33

StataCorp LP. Stata Statistical Software: Release 9. College Station: StataCorp LP; 2005

StataCorp LP. Stata Statistical Software: Release 11. College Station: StataCorp LP; 2009

Steel N, Huppert FA, McWilliams B, Melzer D. Physical and Cognitive Function. In Health, wealth and lifestyles of the older population in England: The 2002 English Longitudinal Study of Ageing. Edited by Marmot M, Banks J, Blundell R, Lessof C, Nazroo J. London: Institute for Fiscal Studies; 2004

Stewart AL, Kamberg CJ. Physical functioning measures. In Measuring Functioning and Well-being: The Medical Outcomes Study approach. Edited by Stewart SL, Ware JE. Durham: Duke Press; 1992

Streiner DL, Norman GR. Health Measurement Scales: a practical guide to their development and use. Oxford: Oxford University Press; 2003

Stromberg U. Collapsing ordered outcome categories: A note of concern. American Journal of Epidemiology 1996; 144: 421-424

Stuck AE, Walthert JM, Nikolaus T, Büla CJ, Hohmann C, Beck JC. Risk Factors for functional status decline in community living elderly people: a systematic literature review. Social Science and Medicine 1999; 48: 445-469

Stucki G, Daltroy L. Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. Journal of Clinical Epidemiology 1996; 49: 711-717

Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. Journal of Rehabilitation Medicine 2001; 33: 47-48

Swanson G, Carrothers L, Mulhorn KA. Comparing disability survey questions in five countries: a study using the ICF to guide comparisons. Disability and Rehabilitation 2003; 25: 665-675

Syddall HE, Martin HJ, Harwood RH, Cooper C, Sayer AA. The SF-36: a simple, effective measure of mobility-disability for epidemiological studies. The Journal of Nutrition, Health and Aging 2009; 13: 57–62

Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value in Health 2004; 7: S22-S26

Tennant A. Goal attainment scaling: Current methodological challenges. Disability and Rehabilitation 2007; 29: 1583-1588

Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis and Rheumatism 2007; 57: 1358-1362

Tennant A, Horton M, Pallant JF. An Introduction to Rasch analysis using RUMM2020. The University of Leeds: Psychometric Laboratory for Health Sciences, Department of Rehabilitation Medicine; 2008

Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 2007; 60: 34-42

Thomas R. Question bank commentary: Income. University of Surrey; 1999 Available at http://surveynet.ac.uk/sqb/ Last accessed 13 September 2010

Thomas E, Wilkie R, Peat G, Hill S, Dziedzic K, Croft P. The North Staffordshire Osteoarthritis Project-NorStOP: Prospective, 3-year study of the epidemiology and management of clinical osteoarthritis in a general population of older adults. BMC Musculoskeletal Disorders 2004a; 5: 2

Thomas E, Peat G, Harris L, Wilkie R, Croft PR. The prevalence of pain and pain interference in a general population of older adults: cross-sectional findings from the North Staffordshire Osteoarthritis Project (NorStOP). Pain. 2004b; 110: 361-8

Thomas R. Measuring disability in the general population using sample surveys and censuses. ESRC Questionbank; 2007. Available at http://qb.soc.surrey.ac.uk/docs/topic_commentaries.htm Last accessed 13 September 2010

Thomas E, Mottram S, Peat G, Wilkie R, Croft P. The effect of age on the onset of pain interference in a general population of older adults: prospective findings from the North Staffordshire Osteoarthritis Project (NorStOP). Pain 2007; 129: 21-27

Tibble, M. Review of Existing Research on the Extra Cost of Disability. Working Paper No. 21. London: Department of Work and Pensions, 2005

Tomassini C. 2005 Demographic profile. In Focus on Older People. Edited by Soule A, Babb P, Evandrou M, Balchin S, Zealey L. London: Stationary Office; 2005: 1-9

Toschke AM, van Kries R, Beyerlein A, Rückinger S. Risk factors for childhood obesity: shift of the entire BMI distribution vs. shift of the upper tail only in a cross sectional study. BMC Public Health 2008; 8: 115

Twisk JWR. Applied longitudinal data analysis for epidemiology. A practical guide. Cambridge: Cambridge University Press; 2003

United Nations. United Nations Statistics Division; 2007 Available at http://unstats.un.org/unsd/demographic/sconcerns/disability/disab2.asp Last accessed 13 September 2010

University of Essex. UK Data Archive; 2007. Available at http://www.data-archive.ac.uk/ last accessed 13 September 2010

University of Surrey. ESRC Question Bank; 2007. Available at http://surveynet.ac.uk/sqb/ Last accessed 13 September 2010

Vach W. The dependence of Cohen's Kappa on the prevalence does not matter. Journal of Clinical Epidemiology 2005; 58: 655-661

van den Brink CL, Tijhuis M, van den Bos, Giampaoli S, Kivinen P, Nissinen A, Kromhout D. effect of widowhood on disability onset in elderly men from three European countries. Journal of the American Geriatrics Society 2004; 52: 353-358

van der Windt DAWM, Dunn KM, Spies-Dorgelo MN, Mallen CD, Blankenstein AH, Stalman WAB. Impact of physical symptoms on perceived health in the community. Journal of Psychosomatic Research 2008; 64: 265-274

Visser M, Langlois J, Guralnik JM, Cauley JA, Kronmal RA, Robbins J, Williamson JD, Harris TB. High nody fatness, but not low fat-free mass, predicts disability in older men and women: the Cardiovascular Health Study. American Journal of Clinical Nutrition 1998; 68: 584-59

Vogt WP. Dictionary of Statistics and Methodology. Newbury Park: Sage; 1993

Wannamethee SG, Ebrahim S, Papacosta O, Shaper AG. From a postal questionnaire of older men, health life style factors reduced the onset of and may have increased recovery from mobility limitation. Journal of Clinical Epidemiology 2005; 58: 831-840

Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care 1992; 30: 473-483

Ware J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Medical Care 1996; 34: 220-233

Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey. Manual and Interpretation Guide. Lincoln Quality Metric; 2000

Ware JE. SF-36 health survey update. Spine 2000; 25: 3130-3139

Weiss CO, Fried LP, Bandeen-Roche K. Exploring the hierarchy of mobility performance in high-functioning older women. Journal of Gerontology. Series A, Biological Sciences and Medical Sciences 2007; 62: 167-173

Wharrad H. Levels of Measurement. Universities' collaboration in eLearning. 2004 Available at: http://www.ucel.ac.uk/showroom/levels_of_measurement/downloads/levels_notes.pdf Last accessed 13 September 2010

Wilkie R, Peat G, Thomas E, Hooper H, Croft PR. The Keele Assessment of Participation: a new instrument to measure participation restriction in population studies. Combined qualitative and quantitative examination of its psychometric properties. Quality of Life Research 2005; 14: 1889-1899.

Williams R. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. Stata Journal 2006; 6:58-82

Winsteps.com. Winsteps computer program. Available at: www.winsteps.com/winsteps.htm last accessed 13 September 2010

Wolinsky FD, Miller TR, Malmstrom TK, Miller JP, Schootman M, Andresen EM, Miller DK. Four-year lower extremity disability trajectories among African American men and women. Journals of Gerontology. Series A, Biological Sciences and Medical Sciences

WHO. International Classification of Impairments, Disabilities and Handicaps. Geneva: World health Organisation; 1980.

WHO. WHODASII disability assessment schedule. Training manual: a guide to administration. Geneva : World Health Organisation; 2000

WHO. International Classification of Functioning, Disability and Health: ICF. Geneva: WHO; 2002

WHO. Global database on Body Mass Index. BMI Classification. 2006 Available at http://apps.who.int/bmi/index.jsp?introPage=intro_3.html Last accessed 13 September 2010

Wright BD, Tennant A. Sample size again. Rasch Measurement Transactions 1996; 9: 468

Wright B, Douglas G. Best test design and self-tailored testing. Mesa memorandum No. 19, University of Chicago: Department of Education; 1975

Wright BD, Stone MH. Best test design: Rasch Measurement. Chicago: Mesa Press;1979

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Archives of Physical Medicine and Rehabilitation 1989; 70: 857-860

Wright BD. Rack and Stack: Time 1 vs. Time 2. Rasch Measurement Transactions 2003; 17: 905-906

Wu ML, Adams RJ, Wilson MR. ACER ConQuest. Melbourne: Acer Press; 1998

Young H, Grundy E, Kalogirou S. Who cares? Geographic variation in unpaid caregiving in England and Wales: evidence from the 2001 Census. Population Trends 2005; 120: 23-34

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandinavica 1983; 67: 361-370

Appendices

Appendix A Search strategy: The measurement of lower limb locomotor disability and its occurrence in middle- and old-age

Table A1 Medline search strategy

	Locomotor disability	Study types	Population	Methods
Terms	1. Mobil$ 2. Limit$ 3. Locomot$ 4. Disab$ 5. Lower extremity 6. (Function or functioning or functions or functioned) 7. (Walk or walker or walking or walks or walked) 8. ambulat$	1. cross sectional 2. longitudinal 3. (epidemiology or epidemiological or epidemiologist) 4. prospective 5. follow up 6. survey 7. health survey 8. cohort	1. Community 2. Dwelling 3. Living 4. Population based	1. Questionnaire 2. (Self report or self reported)
Combinations	(1 NEAR 2) OR (1 NEAR 4) OR (1 NEAR 6) OR (2 NEAR 3) OR (2 NEAR 7) OR (2 NEAR 8) OR (3 NEAR 4) OR (3 NEAR 6) OR (4 NEAR 5) OR (4 NEAR 7) OR (4 NEAR 8) OR (5 NEAR 6) OR (6 NEAR 7) OR (6 NEAR 8)	1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8	(1 ADJ 2) OR (1 ADJ 3) OR 4	1 OR 2

Overall combination: (Population OR Methods) AND Locomotor disability AND Study types

Table A2 EMBASE search strategy

	Locomotor disability	Study types	Population	Methods
Terms	1. (Mobile or mobility or mobilis$ or mobiliz$) 2. (limited or limits or limitation) 3. Locomot$ 4. (Disability$ or disable$) 5. Lower extremity 6. (Function or functioning or functions or functioned) 7. (Walk or walker or walking or walks or walked) 8. Ambulat$ 9. Leg movement	1. Cross sectional 2. Longitudinal 3. Prospective 4. Epidemiol$ 5. Follow up 6. Survey 7. Postal survey 8. Health survey 9. Cohort	1. Community 2. Dwelling 3. Living 4. Population based	1. Questionnaire 2. Self report$
Combinations	(1 NEAR 2) OR (1 NEAR 4) OR (1 NEAR 6) OR (2 NEAR 3) OR (2 NEAR 7) OR (2 NEAR 8) OR (2 NEAR 9)) OR (3 NEAR 4) OR (3 NEAR 6) OR (4 NEAR 5) OR (4 NEAR 7) OR (4 NEAR 8) OR (4 NEAR 9) OR (5 NEAR 6) OR (6 NEAR 7) OR (6 NEAR 8)	1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	(1 ADJ 2) OR (1 ADJ 3) OR 4	1 OR 2

Overall combination: (Population OR Methods) AND Locomotor disability AND Study types

Table A3 DH-Data search strategy

	Locomotor disability	Study types	Population	Methods
Terms	1. Mobil$ 2. Limit$ 3. Locomot$ 4. Disab$ 5. Lower extremity 6. Function$ 7. Walk$ 8. Ambulat$	1. Cross sectional 2. Longitudinal 3. Epidemiol$ 4. Prospective 5. Follow up 6. Survey 7. Cohort	1. Community 2. Dwelling 3. Living 4. Population based	1. Questionnaire 2. Self report$
Combinations	(1 NEAR 2) OR (1 NEAR 4) OR (1 NEAR 6) OR (2 NEAR 3) OR (2 NEAR 7) OR (2 NEAR 8) OR (3 NEAR 4) OR (3 NEAR 6) OR (4 NEAR 5) OR (4 NEAR 7) OR (4 NEAR 8) OR (5 NEAR 6) OR (6 NEAR 7) OR (6 NEAR 8)	1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7	(1 ADJ 2) OR (1 ADJ 3) OR 4	1 OR 2

Overall combination: (Population OR Methods) AND Locomotor disability AND Study types

Appendix B The North Staffordshire Osteoarthritis Project

Appendix B1 Short Physical Performance Battery

The following sections display the data collection form and scoring mechanism for the Short Physical Performance Battery (SPPB) (Guralnik et al 1994).

B1.1 Data collection

The data collection form completed by research therapists during the CAS-HA baseline clinical assessments: pages B2-B3

SHORT PHYSICAL PERFORMANCE BATTERY

***** Remember to check front sheet for red flags, contraindications, and cautions *****

51. Balance Tests

a. Side-by-Side Stand *(tick one)*

Held for 10 sec	☐		**Continue with Semi-Tandem Stand**
Not held for 10 sec	☐	No. of secs held? ☐.☐☐	**Circle a reason; GO TO CHAIR STAND TEST**
Not attempted	☐		**Circle a reason; GO TO CHAIR STAND TEST**

b. Semi-Tandem Stand *(tick one)*

Held for 10 sec	☐		**Continue with Tandem Stand**
Not held for 10 sec	☐	No. of secs held? ☐.☐☐	**Circle a reason; GO TO CHAIR STAND TEST**
Not attempted	☐		**Circle a reason; GO TO CHAIR STAND TEST**

c. Tandem Stand *(tick one)*

Held for 10 sec	☐		
Not held for 10 sec	☐	No. of secs held? ☐.☐☐	**Circle a reason; GO TO CHAIR STAND TEST**
Not attempted	☐		**Circle a reason; GO TO CHAIR STAND TEST**

If participant did not attempt **BALANCE TEST** or failed, circle a reason why:

1	Tried but unable	**5**	Participant unable to understand instructions
2	Participant could not hold position unassisted	**6**	Other (specify) _____
3	Not attempted, you felt unsafe	**7**	Participant refused
4	Not attempted, participant felt unsafe		

Comments

---

52. Chair Stand Tests

If THR/TKR within the last 3 months — **Circle Other (Specify); STOP**

a. Single Chair Stand Test *(tick one)*

Participant stood without using arms	☐	**Continue Repeated Chair Stand Test**
Test not attempted or failed	☐	**Circle a reason; STOP**

b. Repeated Chair Stand Test *(tick one)*

Five stands completed in 60 sec or less	☐	Time taken? ☐☐.☐☐	**STOP**
Test not attempted or failed	☐		**Circle a reason; STOP**

If participant did not attempt **CHAIR STAND TEST** or failed, circle a reason why:

1	Tried but unable	**5**	Participant unable to understand instructions
2	Participant could not stand unassisted	**6**	Other (specify) _____
3	Not attempted, you felt unsafe	**7**	Participant refused
4	Not attempted, participant felt unsafe		

Comments

SHORT PHYSICAL PERFORMANCE BATTERY (cont'd)

***** Remember to check front sheet for red flags, contraindications, and cautions *****

53. Gait Speed Tests

Assessor's initials: ☐☐☐

a. First Gait Speed Test

Aids for first walk? *(tick one)* None ☐
Cane ☐
Other ☐

	(tick one)		
Completed in 60sec or less	☐	Time taken? ☐☐.☐☐	**Continue 2nd Gait Speed Test**
Test not attempted or failed	☐		**Circle a reason; STOP**

b. Second Gait Speed Test

Aids for 2nd walk? *(tick one)* None ☐
Cane ☐
Other ☐

	(tick one)		
Completed in 60sec or less	☐	Time taken? ☐☐.☐☐	**STOP**
Test not attempted or failed	☐		**Circle a reason; STOP**

If participant did not attempt **GAIT SPEED TEST** or failed, circle a reason why:

1 Tried but unable
2 Participant could not walk unassisted
3 Not attempted, you felt unsafe
4 Not attempted, participant felt unsafe
5 Participant unable to understand instructions
6 Other (specify) _____
7 Participant refused

Comments

B1.2 Score formation

The total SPPB score ranges from 0 (poorest ability) to 12 (best ability). This total score formed by summing the score on each of three subscale scores: balance, chair stands and gait speed. The scoring mechanism for each of these subscales is explained below.

Balance

The score for balance was derived as follows from three tasks: side-by-side stand, semi-tandem stand, tandem stand (Table B1).

Table B1 Balance tasks and score allocation

Balance task completed	Points
Side-by-side < 10 seconds	0
Side-by-side only ≥ 10 seconds	1
Both side-by-side & semi-tandem only ≥ 10 seconds	2
Side-by-side & semi-tandem ≥ 10 seconds and tandem ≥ 3 seconds	3
Side-by-side, semi-tandem & tandem ≥ 10 seconds	4

Side-by-side: Participant stands with feet together

Semi-tandem: Participant stands with the side of the heel of one foot touching the big toe of the other foot

Tandem: Participant stands with the heel of one foot in front of and touching the toes of the other foot

Chair stands

The score for chair stands was derived as follows from the time to complete five chair stands (Table B2).

Table B2 Chair standing tasks and score allocation

Time taken to complete five chair stands (seconds)	Points
> 60 or not completed	0
≥ 16.7 & <60	1
≥ 13.7 & <16.7	2
≥ 11.2 & < 13.7	3
< 11.2	4

Gait speed

The score for gait speed was derived from the time taken to walk eight feet (approximately 2.44 metres) (Table B3).

Table B3 Score allocation for time taken to walk eight feet

Time taken to walk eight feet[a] (seconds)	Points
> 60.00 or not completed	0
> 8.70 & ≤ 60.00	1
> 6.20 & ≤ 8.70	2
> 4.81 & ≤ 6.20	3
≤ 4.81	4

[a] Approximately 2.44 metres

Appendix B2 NorStOP Health Survey questionnaire: baseline

The following pages display the Health Survey questionnaire completed by members of the NorStOP cohort at baseline: pages B7-B41.

Health

Questionnaire

Instructions for this questionnaire

The aim of this questionnaire is to find out about the general health of local people and about how many people suffer from joint pain. The answers you give in the questionnaire will be treated in the **strictest confidence.**

Please answer even if you have no problems with your health

Please answer all of the questions unless the instructions ask you to do something else.

Please write in **BLOCK CAPITALS** where appropriate.

Most of the questions can be answered by putting a cross in a box like this: ☒

When you have finished please check that you have answered all of the questions and then return the questionnaire in the envelope enclosed. You do not need a stamp. Please return the questionnaire in the next two weeks.

More details about this project are available in the information sheet enclosed. If you would like further information about this project please contact Ross Wilkie, survey co-ordinator, on 01782 583904.

Thank you for your help with this important research project

Section A

This section is made up of questions about your health, the activities you do, and some of the ways in which people do things in everyday life.

Please answer each set of questions as the instructions tell you to.

Part 1 - Your health

We are interested in your general health.

Please answer every question. Some questions may look similar to others but each one is different. We would like you to answer each one. Please take the time to read and answer each question carefully by placing a cross in the box of your choice. Please cross one box only on each line.

1. In general would you say your health is:
   (Please put a cross in one box only)

Excellent	Very good	Good	Fair	Poor
☐	☐	☐	☐	☐

2. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?
   (Please put a cross in one box on each line)

	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf.....	☐	☐	☐
b. Climbing **several** flights of stairs.....	☐	☐	☐

3. During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities <u>as a result of your physical health</u>?
   (Please put a cross in one box on each line)

	Yes	No
a. **Accomplished less** than you would like.....	☐	☐
b. Were limited in the **kind** of work or other activities	☐	☐

4. During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?
(Please put a cross in one box on each line)

	Yes	No
a) **Accomplished less** than you would like	☐	☐
b) Didn't do work or other activities as **carefully** as usual	☐	☐

5. During the **past 4 weeks**, how much did **pain** interfere with your normal work (including both work outside the home and housework)?
(Please put a cross in one box only)

Not at all	A little bit	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

6. These questions are about how you feel and how things have been with you during the **past 4 weeks**. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the **past 4 weeks**.....
(Please put a cross in one box on each line)

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
a) Have you felt calm and peaceful?.....	☐	☐	☐	☐	☐	☐
b) Did you have a lot of energy?.....	☐	☐	☐	☐	☐	☐
c) Have you felt downhearted and depressed?.....	☐	☐	☐	☐	☐	☐

7. During the **past 4 weeks,** how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)? **(*Please put a cross in one box only)***

All of the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

Part 2 - Taking Part

We are interested in some of the things that are necessary for you to live your life in the way you choose. We are particularly interested in how often these things are achieved in the way you would like.

When answering the questions, please think about the **past 4 weeks.** It does not matter if you require the help of other people or from gadgets and machines. We would simply like to know if the activity **IS** achieved to the extent that you want it to be.

Please read each statement below and put a cross in the box which comes closest to how much you agree with the statement. Please put a cross in one box only for each line.

1. During the past 4 weeks, I have moved around in my home, **as and when I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

2. During the past 4 weeks, I have moved around outside my home, **as and when I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

3. During the past 4 weeks, my self-care needs (examples are washing, toileting, dressing, feeding, maintaining health) have been met, **as and when I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

4. During the past 4 weeks, my home has been looked after, **as and when I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

5. During the past 4 weeks, my things (belongings) have been looked after, **as and when I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

6. Do you have any relatives, or other people, who depend on you?

Yes.... ☐ No.... ☐

If yes, during the past 4 weeks, were these people looked after, **as and when you have wanted**?

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

7. During the past 4 weeks, I have met and spoken to other people, **as and when I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

8. During the past 4 weeks, I, or someone else on my behalf, have managed my money, **as I have wanted**.

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

9. Do you choose to take part in paid or voluntary work?

Yes.... ☐ No.... ☐

If yes, during the past 4 weeks, have you taken part in paid or voluntary work, **as and when you have wanted**?

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

10. Do you choose to take part in education or training courses?

Yes.... ☐ No.... ☐

If yes, during the past 4 weeks, have you taken part in education or training, **as and when you have wanted**?

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

11. Do you choose to take part in social activities?
(Examples of social activities are community and religious activities, meeting up with friends, going to clubs)

Yes.... ☐ No.... ☐

If yes, during the past 4 weeks, have you taken part in social activities, **as and when you have wanted**?

All the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

Part 3 - How often

We would like to know about how often you do things during your normal daily routine.

Look at the following list. **Thinking about the past 4 weeks**, please consider how often you did each thing, and put a cross in one box on each line.

	All days	Most days	Some days	Few days	No days
a. Go out for a walk.....	☐	☐	☐	☐	☐
b. Go out of the house and go somewhere.....	☐	☐	☐	☐	☐
c. Go out to work.....	☐	☐	☐	☐	☐
d. Spend most or all of the day in bed or in a chair.....	☐	☐	☐	☐	☐
e. Go shopping.....	☐	☐	☐	☐	☐
f. Take a bus or drive a car.....	☐	☐	☐	☐	☐
g. Go in a car as a passenger.....	☐	☐	☐	☐	☐
h. Go to a club, church or social event.....	☐	☐	☐	☐	☐
i. Play a sport.....	☐	☐	☐	☐	☐
j. Go on an education or training course.....	☐	☐	☐	☐	☐
k. Do a hobby.....	☐	☐	☐	☐	☐
l. Take a bath/shower.....	☐	☐	☐	☐	☐
m. Do home maintenance activities.....	☐	☐	☐	☐	☐

	All days	Most days	Some days	Few days	No days
n. Wash clothes.....	☐	☐	☐	☐	☐
o. Cook and clean.....	☐	☐	☐	☐	☐
p. Look after others.....	☐	☐	☐	☐	☐
q. Heavy housework (e.g. spring cleaning, moving furniture, scrubbing floors by hand).....	☐	☐	☐	☐	☐
r. Heavy gardening (e.g.digging, tree felling).....	☐	☐	☐	☐	☐
s. Heavy DIY work at home (e.g. decorating, plastering)..	☐	☐	☐	☐	☐
t. Walks of at least a quarter of a mile (5-10 minutes continuous walking)..	☐	☐	☐	☐	☐
u. Walks of two miles or more (at least 40 minutes continuous walking).....	☐	☐	☐	☐	☐

Part 4 - How you do things

We would like to know if you have changed the way you do your normal daily activities and if you require the help of other people or special devices.

Please answer each question and put a cross in one box for each one.

1. Thinking back over the **past 4 weeks,** have you had to reduce the amount of time or change how you have done most activities because of your health?

All days	Most days	Some days	Few days	No days
☐	☐	☐	☐	☐

2. During the **past 4 weeks** have you required the assistance of others or aids (examples are a walking stick or a wheelchair) to move around your home?

Yes... ☐ No... ☐

3. During the **past 4 weeks**, have you required the assistance of others or aids (examples are a walking stick or a wheelchair) to go to places outside of your home?

Yes... ☐ No... ☐

4. During the **past 4 weeks** have you required the assistance of others to look after yourself?

Yes... ☐ No... ☐

5. During the **past 4 weeks,** have you required the assistance of others to look after your home?

Yes... ☐ No... ☐

6. Compared to **12 months ago,** have you reduced the time or changed how you have done any of your activities?

Yes, a lot	Yes, a little	No, not at all
☐	☐	☐

Part 5 - Friends and Family

We are interested in the contact you may have with your friends and family.
Please answer each question and put a cross in one box for each line.

1. How often do you go to religious meetings or services?

More than once a week	Once a week	1 to 3 times per month	Less than once per month	Never or almost never
☐	☐	☐	☐	☐

2. How many hours **each week** do you participate in any groups such as social or work group, church-connected group, self-help group, charity, public service or community group?

None	1 to 2 hours	3 to 5 hours	6 to 10 hours	11 to15 hours	16 or more hours
☐	☐	☐	☐	☐	☐

3. How many living children do you have?

None	1 to 2	3 to 5	6 or more
☐	☐	☐	☐

4. **How many of your children** do you see at least **once a month**?

None	1 to 2	3 to 5	6 or more
☐	☐	☐	☐

5. Apart from your children, how many relatives do you have with whom you feel close?

None	1 to 2	3 to 5	6 to 9	10 or more
☐	☐	☐	☐	☐

6. **How many close relatives** do you see at least **once a month**?

None	1 to 2	3 to 5	6 to 9	10 or more
☐	☐	☐	☐	☐

7. How many close friends do you have?

None	1 to 2	3 to 5	6 to 9	10 or more
☐	☐	☐	☐	☐

8. **How many of these friends** do you see at least **once a month**?

None	1 to 2	3 to 5	6 to 9	10 or more
☐	☐	☐	☐	☐

9. Is there any one special person you know that you feel very close to; someone you feel you can share confidences and feelings with?

Yes.... ☐ No.... ☐

If yes, how often do you see or talk with this person?

Daily	Weekly	Monthly	Several times a year	Once a year or less
☐	☐	☐	☐	☐

Part 6 - Access

We would like you to tell us how easy it is to get hold of or gain access to things that you need in daily life.

(Please put a cross in one box for each question)

1. Do you have access to a car when you personally need it?

Yes... ☐ No... ☐

2. Do you have access to public transport?

Yes... ☐ No... ☐

3. Do you have access to a telephone?

Yes... ☐ No... ☐

4. Do you have good access to your doctor (GP), as and when you need it?

Yes... ☐ No... ☐

5. Do you have access to a chemist?

Yes... ☐ No... ☐

6. Do you have access to a bank?

Yes... ☐ No... ☐

7. Do you have access to advice or help with your income (for example, relatives or the benefits system)?

Yes... ☐ No... ☐

8. If you wanted to take part in an education course, is there the opportunity?

Yes... ☐ No... ☐

9. If you wanted to do paid or voluntary work, is there the opportunity?

Yes... ☐ No... ☐

Part 7 - How you feel

The next set of questions are about how you feel at the moment. Please read each item and put a cross next to the reply that comes closest to how you have been feeling **in the past week.** Don't take too long over your replies; your immediate reaction to each item will usually be more accurate than a long thought out response.

1. I feel tense or 'wound up':

Most of the time	A lot of the time	From time to time, occasionally	Not at all
☐	☐	☐	☐

2. I still enjoy the things I used to enjoy:

Definitely as much	Not quite as much	Only a little	Hardly at all
☐	☐	☐	☐

3. I get a sort of frightened feeling as if something awful is about to happen:

Very definitely and quite badly	Yes, but not too badly	A little, but it doesn't worry me	Not at all
☐	☐	☐	☐

4. I can laugh and see the funny side of things:

As much as I always could	Not quite so much now	Definitely not so much now	Not at all
☐	☐	☐	☐

5. Worrying thoughts go through my mind:

A great deal of the time	A lot of the time	Not too often	Very little
☐	☐	☐	☐

6. I feel cheerful:

Never	Not often	Sometimes	Most of the time
☐	☐	☐	☐

7. I can sit at ease and feel relaxed:

Definitely	Usually	Not often	Not at all
☐	☐	☐	☐

8. I feel as if I am slowed down:

Nearly all the time	Very often	Sometimes	Not at all
☐	☐	☐	☐

9. I get a sort of frightened feeling like 'butterflies' in my stomach:

Not at all	Occasionally	Quite often	Very often
☐	☐	☐	☐

10. I have lost interest in my appearance:

Definitely	I don't take as much care as I should	I may not take quite as much care	I take just as much care as ever
☐	☐	☐	☐

11. I feel restless as if I have to be on the move:

Very much indeed	Quite a lot	Not very much	Not at all
☐	☐	☐	☐

12. I look forward with enjoyment to things:

As much as I ever did	Rather less than I used to	Definitely less than I used to	Hardly at all
☐	☐	☐	☐

13. I get sudden feelings of panic:

Very often indeed	Quite often	Not very often	Not at all
☐	☐	☐	☐

14. I can enjoy a good book or radio or television programme:

Often	Sometimes	Not often	Very seldom
☐	☐	☐	☐

Part 8- Your views

Section 1

We would like to know what your views are on **what causes illness.**
Please indicate your agreement or disagreement with each of the following as a possible cause of illness.
(Please put a cross in one box on each line)

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
a. Stress or worry.....	☐	☐	☐	☐	☐
b. Hereditary - it runs in the family.....	☐	☐	☐	☐	☐
c. A germ or virus.....	☐	☐	☐	☐	☐
d. Diet or eating habit.....	☐	☐	☐	☐	☐
e. Chance or bad luck.....	☐	☐	☐	☐	☐
f. Poor medical care in the past.....	☐	☐	☐	☐	☐
g. Pollution in the environment.....	☐	☐	☐	☐	☐
h. Someone's own behaviour.	☐	☐	☐	☐	☐
i. Someone's mental attitude (e.g. thinking about life negatively).....	☐	☐	☐	☐	☐
j. Family problems or worries.....	☐	☐	☐	☐	☐
k. Overwork.....	☐	☐	☐	☐	☐
l. Someone's emotional state (e.g. feeling down, lonely, anxious, empty).....	☐	☐	☐	☐	☐
m. Ageing.....	☐	☐	☐	☐	☐

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
n. Alcohol.....	☐	☐	☐	☐	☐
o. Smoking.....	☐	☐	☐	☐	☐
p. Accident or injury.....	☐	☐	☐	☐	☐
q. Someone's personality.....	☐	☐	☐	☐	☐
r. Altered immunity (the body is unable to fight illness).....	☐	☐	☐	☐	☐

Section 2

We are also interested in how you think illness can be controlled and how you may use your doctor (GP).

1. In your opinion, is it a matter of luck whether you are well or ill, or is it something which can be controlled?
   (Please put a cross in one box only)

All luck	Mostly luck	Bit of both	Mostly under control	Almost all under control
☐	☐	☐	☐	☐

2. How often do you visit the doctor (GP) for yourself?
   (Please put a cross in one box only)

Very often	Often	Occasionally	Seldom	Hardly ever
☐	☐	☐	☐	☐

3. When you are ill, do you usually go straight away to the doctor, or do you wait to see if you get better? ***(Please put a cross in one box only)***

a. Go straight to the doctor..... ☐

b. Wait a day or two to see if it gets better..... ☐

c. Wait several days to see if it gets better..... ☐

d. Put it off as long as possible..... ☐

Section 3

We are interested in your views of health and life. Please indicate how much you agree or disagree with the following statements about health issues by putting a cross in one box for each statement.

(Please put a cross in one box on each line)

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
a. There is a lot which I can do to control my health.....	☐	☐	☐	☐	☐
b. What I do will affect whether my health gets better or worse.....	☐	☐	☐	☐	☐
c. Treatments are effective in controlling illness.....	☐	☐	☐	☐	☐
d. My health is very unpredictable.....	☐	☐	☐	☐	☐
e. Illness makes me feel afraid.....	☐	☐	☐	☐	☐
f. The course of my life depends on me.....	☐	☐	☐	☐	☐
g. I have the power to influence what happens in my life.....	☐	☐	☐	☐	☐
n. Osteoarthritis is a serious condition.....	☐	☐	☐	☐	☐

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
i. Problems with your joints are a normal part of growing old.....	☐	☐	☐	☐	☐
j. Doctors can do a lot to help people with joint problems.....	☐	☐	☐	☐	☐
k. Joint problems always get worse over time.....	☐	☐	☐	☐	☐
Treatments are effective in controlling pain.....	☐	☐	☐	☐	☐
m. If a pain lasts for a week or more, you have a serious illness.....	☐	☐	☐	☐	☐
n. I do not expect doctors to be able to do much about pain.....	☐	☐	☐	☐	☐
o. The thought of pain makes me afraid.....	☐	☐	☐	☐	☐

Part 9 - About you

Here are some general questions about yourself.

Please follow the instructions and answer ALL of the following questions.

1. What is your date of birth? ☐☐ / ☐☐ / ☐☐

(For example - if you were born on the 5th June 1936, this would be entered as 05/06/36)

2. Are you:

Female..... ☐ Male..... ☐

3. What is your current marital status? ***(Please put a cross in one box only)***

Married..... ☐ Widowed..... ☐

Separated..... ☐ Cohabiting..... ☐

Divorced..... ☐ Single..... ☐

4. Do you live alone?

Yes..... ☐ No..... ☐

5. What is your current employment status?
(Please put a cross in one box only)

Employed..... ☐

Not working due to ill health..... ☐

Retired..... ☐

Unemployed/seeking work..... ☐

Housewife..... ☐

Other..... ☐

6. If working, what is your job title (examples - factory worker, welder, office worker, shop assistant, lawyer) ?

.....

If you are not working, or are retired, what was your last job title?

.....

7. Do you have a spouse or partner who is currently living with you?

Yes... ☐ No... ☐

If they are working, what is their job title (examples - factory worker, welder, office worker, shop assistant, lawyer)?

.....

If they are not working or are retired, what was their last job title?

.....

If you are a widow, or widower, what was your spouse's last job title?

.....

8. What is your weight? ☐☐ stones ☐☐ lbs or ☐☐☐ kgs

9. What is your height? ☐☐ feet ☐☐ inches or ☐☐☐ cms

10. What is your current smoking status?

(Please put a cross in one box only)

Never smoked..... ☐

Previously smoked..... ☐

Currently smoking..... ☐

11. On average, how often do you drink alcohol?
(Please put a cross in one box only)

Daily or most days	Once or twice a week	Once or twice a month	Once or twice a year	Never
☐	☐	☐	☐	☐

12. How old were you when you left school? ☐☐ years old

13. Did you go from school to full-time education or university?

Yes..... ☐

No..... ☐

If yes, what age did you finish full-time education? ☐☐ years old

14. Have you gained qualifications through study as an adult?

Yes..... ☐

No..... ☐

15. Thinking about the cost of living as it affects you, which of these descriptions best describes your situation:
(Please put a cross in one box only)

Find it a strain to get by from week to week..... ☐

Have to be careful with money..... ☐

Able to manage without much difficulty..... ☐

Quite comfortably off..... ☐

16. Is your ethnic origin? ***(Please put a cross in one box only)***

White UK/European..... ☐

Afro Caribbean..... ☐

Chinese..... ☐

Asian..... ☐

African..... ☐

Other..... ☐

Part 10 - About your health

We would like to know if you have any other health problems. Please put a cross in the box if you suffer from any of the listed problems.

1. Do you suffer from any of the following?
   (Please place a cross in the box for any which apply to you)

a. Chest problems..... ☐

b. Heart problems..... ☐

c. Deafness..... ☐

d. Problems with eyesight (excluding the need for glasses)..... ☐

e. Raised blood pressure..... ☐

f. Diabetes..... ☐

2. Thinking back over **the past 3 months**, have you suffered from any of the following?
   (Please place a cross in the box for any which apply to you)

a. A fall or falls..... ☐

b. Difficulty remembering things..... ☐

c. Cough with spit..... ☐

d. Breathless when walking..... ☐

e. Dizziness or unsteadiness..... ☐

f. Weakness in an arm or leg..... ☐

3. Please put a cross in the box to show whether you agree (yes box) or disagree (no box) with each of the following statements.....

(Please put a cross in one box on each line)

a. I am confused and start to do more than one thing at a time.....Yes.... ☐ No.... ☐

b. I have more minor accidents than usual (e.g.I drop things, I trip and fall, or bump into things).....Yes.... ☐ No.... ☐

c. I react slowly to things that are said or done.....Yes.... ☐ No.... ☐

d. I do not finish things that I start.....Yes.... ☐ No.... ☐

e. I have difficulty reasoning and solving problems (e.g. making plans, making decisions, or learning new things).....Yes.... ☐ No.... ☐

f. I sometimes get confused (e.g. I do not know where I am, who is around, or what day it is).....Yes.... ☐ No.... ☐

g. I forget a lot (e.g. things that happened recently, where I put things, or to keep appointments).....Yes.... ☐ No.... ☐

h. I do not keep my attention on any activity for long.....Yes.... ☐ No.... ☐

i. I make mistakes more than usual.....Yes.... ☐ No.... ☐

j. I have difficulty doing things which involve thought and concentration..... Yes.... ☐ No.... ☐

4. Thinking over the **past 4 weeks,** did you?

(Please put a cross in one box on each line)

	Not at all	On some nights	On most nights
a. Have trouble falling asleep.....	☐	☐	☐
b. Wake up several times per night.....	☐	☐	☐
c. Have trouble staying asleep.....	☐	☐	☐
d. Wake up after your usual amount of sleep feeling tired and worn out.....	☐	☐	☐

Thank you for completing Section A

Please turn to Section B

Section B

This section is about any pains you may have or any problems with your joints.

Please fill in all the parts even if you do not suffer from any pains or joint problems.

Please follow the instructions for each part.

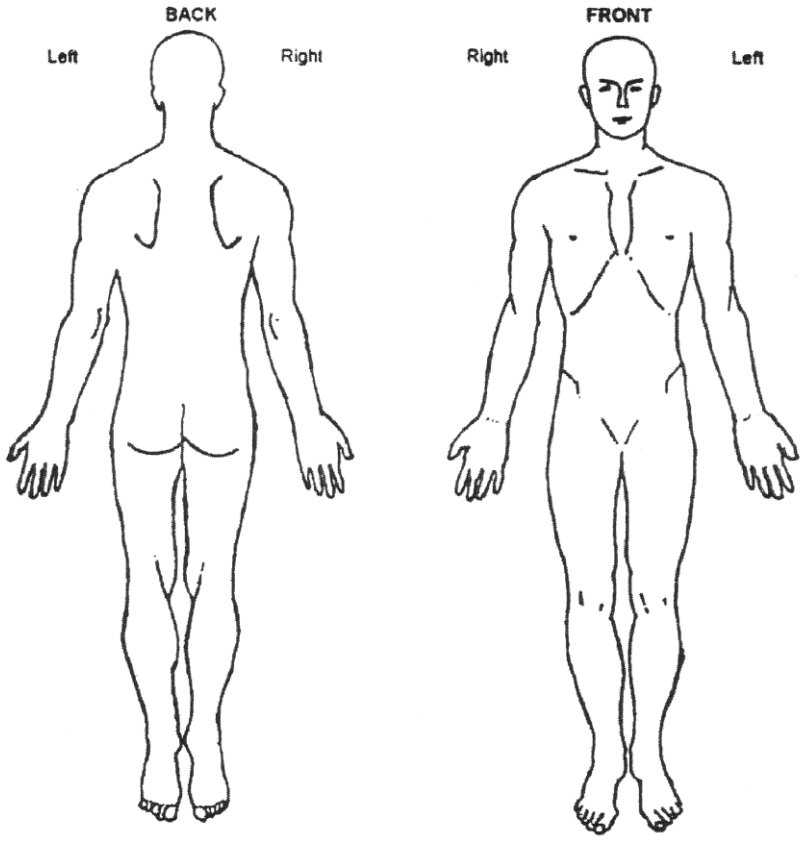
Part 1 - Body Chart

This question is about any **recent pain** you may have had in **any part of your body**. By pain we also mean ache, discomfort or stiffness. Please **do not** include pain due to a feverish illness such as flu. If you are a woman please **do not** include pain related to your monthly period.

1. In the **past 4 weeks**, have you had pain that has lasted for **one day or longer** in **any part of your body**?

Yes ..... ☐ Please shade in the diagram below **any pain** that has lasted for **one day or longer in the past 4 weeks**

No ..... ☐ Please turn the page over and continue with **Part 2.**



Please turn over to complete Part 2

Part 2

It would be helpful if you could tell us about any tablets, pills or creams you may have used to reduce pain, in the **past 4 weeks.**

1. In the **past 4 weeks,** how often have you taken any medicines for your recent pain? ***(Please put a cross in one box on each line)***

	All days	Most days	Some days	Few days	No days
a. Painkillers, Anti-inflammatories (e.g paracetamol, aspirin, Diclofenac,Ibuprofen).....	☐	☐	☐	☐	☐
b. Creams, gels or rub-ons (e.g. Ibuleve, Ibuprofen, Ralgex).....	☐	☐	☐	☐	☐
c. Natural medicine(e.g. herbal remedies, cod liver oil).....	☐	☐	☐	☐	☐
d. Glucosamine/Chondroitin Sulphate.....	☐	☐	☐	☐	☐

Part 3

We are interested to know how many people suffer from specific joint pains and problems. Please answer each of the following questions about your hands, hips, knees, and feet, even if you do not suffer from any problems.

We would like you to tell us if you have had **any problems** with your **hands. 'Hand problems'** relate to **any hand symptoms** you may have experienced; for example, **pain or stiffness,** or any **knobbly swellings** on your hands, including your fingers and thumbs.

1. Have you had any problems with your **HANDS,** including your fingers and thumbs, over the **last year?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

2. Have you had any pain in your **HANDS,** including in your fingers and thumbs, over the **last year?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

We are also interested in **any problems** you may have had with your **legs.** For these questions, please think about problems with your **hip, knee or foot**; for example **pain, stiffness, giving way or locking.**

3. Have you had any problems with your **KNEES**, over the **last year?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

4. Have you had pain in the last year in and **around the KNEE?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

5. Have you had any problems with your **HIPS**, over the **last year?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

6. Have you had pain in the last year in and **around the HIP?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

7. Have you had any problems with your **FEET**, over the **last year?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

8. Have you had pain in the last year in and **around the FOOT?**
   (Please put a cross in one box only)

   Yes.... ☐ No.... ☐

Part 4

The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities? If so, how much?
(Please put a cross in one box on each line)

	Yes, limited a lot	Yes, limited a little	No,not limited at all
a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	☐	☐	☐
b. **Moderate activities,** such as moving a table, pushing a vacuum cleaner, bowling or playing golf	☐	☐	☐
c. Lifting or carrying groceries	☐	☐	☐
d. Climbing **several** flights of stairs	☐	☐	☐
e. Climbing **one** flight of stairs	☐	☐	☐
f. Bending, kneeling, or stooping	☐	☐	☐
g. Walking **more than a mile**	☐	☐	☐
h. Walking **half a mile**	☐	☐	☐
i. Walking **100 yards**	☐	☐	☐
j. Bathing or dressing yourself	☐	☐	☐

Thank you for completing Section B

Please turn to Section C

SECTION C - CONTINUING TO HELP WITH THIS STUDY

Thank you very much for completing this questionnaire.

There will be more stages to this study, and we hope that people who have taken part will be able to help us again.

Please complete the two questions on the opposite page with regards to further help.

Even if you have decided not to participate further in this study or would prefer us not to review your records, the answers in this questionnaire will still be very useful to us

Please check that you have answered all the questions.

Please return this questionnaire in the envelope provided (no stamp needed).

If you have any questions, telephone Ross Wilkie on 01782 583904.

Would you be willing to be contacted again about the possibility of taking part in this study? Giving us your permission to contact you again does not mean that you must take part. You are just agreeing to be contacted again.

Yes, I am happy to be contacted again..... ☐

No, I do not want to be contacted again..... ☐

It is important for us to find out what types of treatments and tests people need. we can do this by reviewing medical records.

Would you be willing to give your permission for this? When we review the medical records, your name will not be used so that you will not be identified personally. We can assure you that any information will be held in the strictest confidence.

Yes, I give my permission for my medical records to be reviewed... ☐

No, I do not want my medical records to be reviewed..... ☐

Signed ..... Date .....

Please print your name and address - .....

.....

.....

.....

Thank you for taking the time to fill in this questionnaire.

SECTION D: CONTINUING TO HELP WITH THIS STUDY

Thank you very much for completing this questionnaire.

There will be further stages to this study, and we hope that people who have taken part so far will be able to help us again.

We may want to contact you again as part of this study and we are asking your permission to do this. Giving us permission to contact you again does not mean that you must take part.

Please put a cross in one of the boxes below to tell us if you are happy to be contacted again, and then sign and date where shown.

Yes, I am happy to be contacted again..... ☐

No, I do not want to be contacted again..... ☐

Please return this questionnaire whether or not you want to be contacted again.

Signature..... Date.....

Please print your **name** and **address** - .....

.....

.....

.....

Please check that you have answered all the questions.

Please return this questionnaire in the envelope provided (no stamp needed).

If you have any questions, telephone Rosie Lacey on 01782 584721.

Thank you for your help with this research project

Appendix B3 Use of the pain screening item and the body manikin for use in Chapter 14

B3.1 Background

There are several different ways in which the screening item and the body manikin could have been combined. Three of these options were considered and compared.

Option1: Respondents must respond “Yes” to the screening item and shade the body manikin, or “No” to the screening item and not shade the body manikin. Those with other patterns of response are excluded.

Option 2: Respondents who shade the manikin are assumed to have pain in the shaded areas. Respondents who do not shade the body manikin are assumed not to have pain. This option makes no use of the screening item.

Option 3: Respondents who shade the manikin are assumed to have pain in the shaded areas. Respondents who do not shade the body manikin and respond ‘No’ to the screening item are assumed not to have pain. Those respondents who do not shade the body manikin and respond ‘Yes’ to the screening item or leave it blank are excluded.

Option 1 is the approach adopted in Chapter 8. This resulted in a high level of missing pain data.

Option 2 is straight forward, but risks allocating people to the ‘no pain’ group because they missed out the body manikin, either because they did not wish to complete it, did not understand how to complete it, or turned two pages together. If the screening item was also incomplete, it was not possible in the baseline questionnaire to ascertain if this final possibility had occurred because the manikin and screening item were on a single page with no other items and the opposite page was blank (Baseline Health Survey questionnaire: Appendix B2: pages 26 and 27).

Option 3 is more complicated to administer, but is attractive in that it does not make assumptions about the meaning of the blank manikins. It does however; make full use of the shaded manikins.

This Appendix describes how a choice was made as to the most suitable option to pursue to define pain for use in Chapter 14.

B3.2 Methods

In order to test empirically which of Options 1, 2 or 3 is the most sensible in these data, a series of analyses were conducted in the NorStOP1, 2 and 3 baseline dataset (n=18,497). The proportion of people in each of the six pain categories, as used in Chapter 8 was calculated under Options 1, 2 and 3. Comparisons were made by calculating the ratio of the proportions in each of the six pain categories: Option 1 / Option 2; Option 1 / Option 3. The option that produced the most ratios closest to 1 in comparison with Option 1 was considered the most appropriate.

B3.3 Results

Table B4 shows the proportions of people in each pain category (comparisons: Option 1 / Option 2; Option 1 / Option 3).

Treating Option1 as the 'gold standard', Option 2 tended to over-estimate the proportion of people who did not have body pain, whilst Option 3 tended to underestimate the proportion with no pain. Whilst Option 3 tended to deviate from Option 1 more often (arbitrarily considering a range of 0.95 to 1.05 for the ratio of percentages to be acceptable), Option 2, when it deviated from Option 1 did so more widely. This was especially true in those people with pain interference and specific joint pains.

Table B4 Proportions of people respondents in each pain category under Options 1, 2 and 3

	Option 1, n(%)	Option 2, n(%)	Option 3, n(%)	Proportion ratio: Option 1 / Option 2	Proportion ratio: Option 1 / Option 3
No pain	4745 (29.0)	5758 (31.1)	4745 (27.1)	0.93	1.07
LLP only	1515 (9.3)	1706 (9.2)	1706 (9.8)	1.00	0.95
LLP & EP	2481 (15.2)	2721 (14.7)	2721 (15.6)	1.03	0.97
LLP & LBP	1513 (9.3)	1657 (9.0)	1657 (9.5)	1.03	0.98
LLP & EP & LBP	4260 (26.0)	4573 (24.7)	4573 (26.2)	1.05	1.00
Pain, not LLP	1844 (11.3)	2082 (11.3)	2081 (11.9)	1.00	0.96

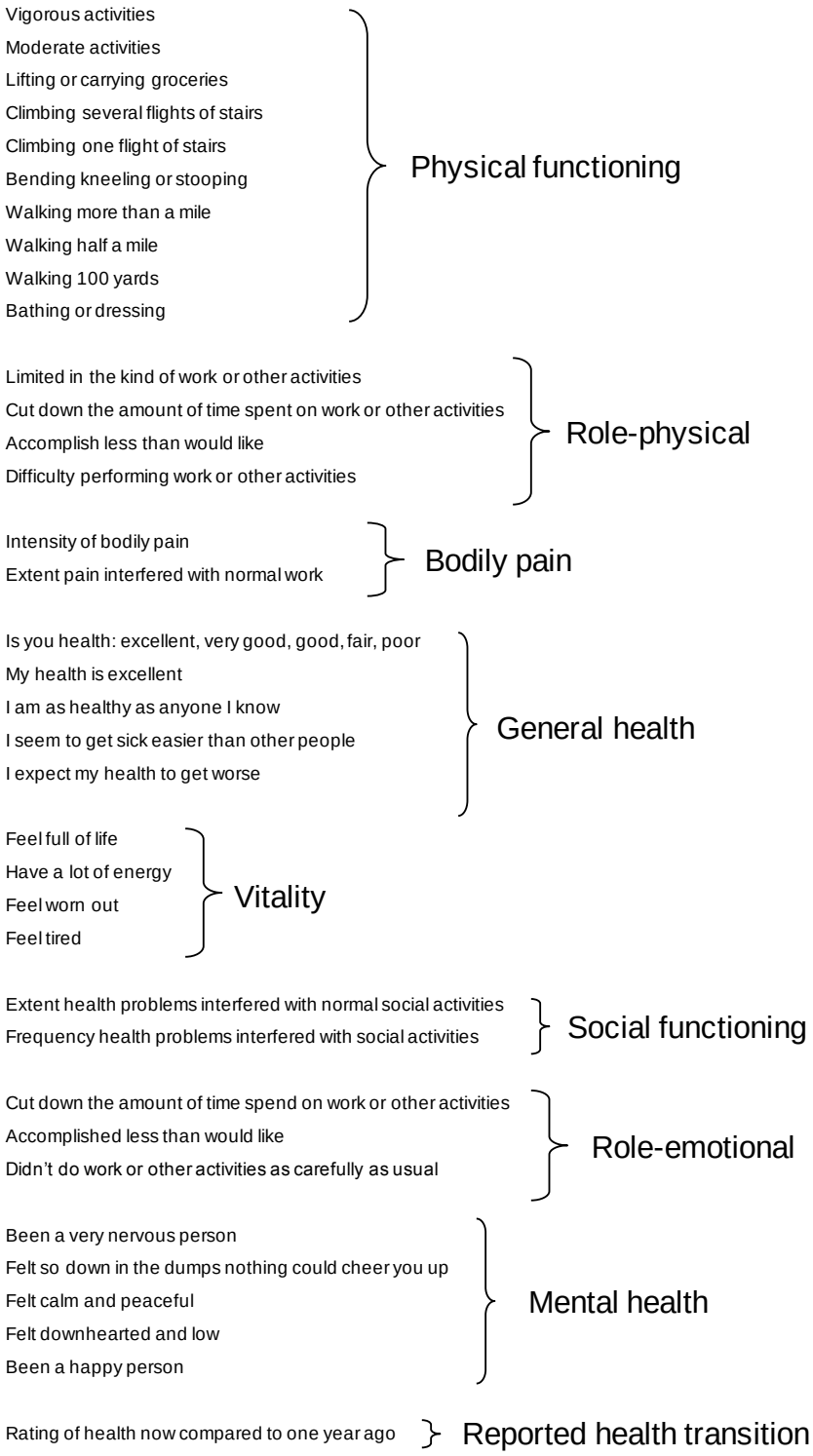
LLP – lower limb pain; LBP – low back pain; EP – elsewhere pain, i.e. not in lower limb or low back

B3.4 Conclusions

Due to the sometimes-large deviations from Option 1 when using Option2, especially in those with pain, it was decided to use Option 3 in further analyses. Although this may result in a more general deviation from Option 1, than would the choice of using Option 2 in further analyses, it will ensure that this deviation is fairly constant. Should Option 2 be used, it would be more likely to underestimate associations between pain and locomotor disability (one of the main interests in Chapter 14), because there may be people assigned to the 'no pain' group, who in actual fact did have pain. Option 3 more closely guards against this, whilst still increasing the numbers available for analysis above the numbers available under Option 1.

Appendix C The suitability of the SF-36 Physical Functioning subcale (PF-10) as a measure of locomotor disability

Appendix C1 Structure of the SF-36: 36 items mapping to eight subscales



Appendix C2 Results of repeatability testing of individual locomotor disability-specific PF-10 items

Table C1 Agreement between individual PF-10 items in those who did not report a change in general health. NorStOP pilot study, test-retest component, n=131

d (climb several flights of stairs)		Retest			
		Limited a lot	Limited a little	Not limited at all	Total
Test	Limited a lot	33	3	1	37
	Limited a little	7	33	5	45
	Not limited at all	0	6	42	48
	Total	40	42	48	130

e (climb one flight of stairs)		Retest			
		Limited a lot	Limited a little	Not limited at all	Total
Test	Limited a lot	10	3	0	13
	Limited a little	2	25	5	32
	Not limited at all	0	5	80	85
	Total	12	33	85	130

g (walk more than a mile)		Retest			
		Limited a lot	Limited a little	Not limited at all	Total
Test	Limited a lot	24	10	1	35
	Limited a little	6	19	5	30
	Not limited at all	1	6	59	66
	Total	31	35	65	131

h (walk half a mile)		Retest			
		Limited a lot	Limited a little	Not limited at all	Total
Test	Limited a lot	20	7	1	28
	Limited a little	7	16	4	27
	Not limited at all	1	8	63	72
	Total	28	31	68	127

i (walk 100 yards)		Retest			
		Limited a lot	Limited a little	Not limited at all	Total
Test	Limited a lot	11	3	0	14
	Limited a little	2	18	4	24
	Not limited at all	0	5	86	91

Total	13	26	90	129

Appendix C3 Results of construct validity testing of individual locomotor disability-specific PF-10 items

Table C2 Hypotheses 1a (chest problems): Results

Item	Chest problems	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	No	27.1	32.6	40.3
	Yes	58.4	26.9	14.7
e (climb one flight of stairs)	No	12.9	23.0	64.1
	Yes	32.8	35.5	31.7
g (walk more than a mile)	No	27.3	24.4	48.3
	Yes	55.6	21.8	22.7
h (walk half a mile)	No	20.0	17.1	62.9
	Yes	44.4	21.4	34.2
i (walk 100 yards)	No	9.8	15.8	74.4
	Yes	25.9	26.6	47.5

Table C3 Hypotheses 1b (heart problems): Results

Item	Heart problems	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	No	27.7	32.8	39.5
	Yes	62.1	24.9	13.0
e (climb one flight of stairs)	No	13.8	23.1	63.1
	Yes	32.7	37.7	29.7
g (walk more than a mile)	No	27.9	24.4	47.7
	Yes	58.5	21.4	20.1
h (walk half a mile)	No	20.6	17.2	62.2
	Yes	47.0	21.7	31.3
i (walk 100 yards)	No	10.7	15.7	73.7
	Yes	25.2	29.7	45.2

Table C4 Hypotheses 1c (falls): Results

Item	Falls	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	No	29.9	32.8	37.3
	Yes	60.8	21.7	17.6
e (climb one flight of stairs)	No	13.9	24.9	61.3
	Yes	39.8	31.2	29.1
g (walk more than a mile)	No	29.2	25.0	45.9
	Yes	62.8	16.1	21.1
h (walk half a mile)	No	21.2	18.2	60.6
	Yes	53.7	16.5	29.8
i (walk 100 yards)	No	10.4	16.9	72.7
	Yes	33.3	26.7	40.0

Table C5 Hypotheses 1d (breathlessness when walking): Results

Item	Breathlessness when walking	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	No	19.5	31.2	49.4
	Yes	59.8	31.8	8.4
e (climb one flight of stairs)	No	9.6	17.6	72.8
	Yes	30.9	40.4	28.7
g (walk more than a mile)	No	20.7	22.2	57.2
	Yes	56.4	26.9	16.7
h (walk half a mile)	No	15.0	13.9	71.1
	Yes	43.9	25.4	30.6
i (walk 100 yards)	No	7.7	11.4	80.9
	Yes	23.5	30.4	46.2

Table C6 Hypotheses 1e (dizziness or unsteadiness): Results

Item	Dizziness or unsteadiness	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	No	25.5	33.1	41.5
	Yes	59.6	26.3	14.1
e (climb one flight of stairs)	No	11.5	22.8	65.8
	Yes	34.7	34.6	30.7
g (walk more than a mile)	No	25.4	24.5	50.1
	Yes	58.1	21.8	20.0
h (walk half a mile)	No	18.2	17.0	64.8
	Yes	47.1	21.2	31.7
i (walk 100 yards)	No	8.8	15.2	76.1
	Yes	27.1	27.4	45.5

Table C7 Hypotheses 1f (high BMI): Results

Item	BMI group	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Normal	28.8	30.1	41.1
	High	35.6	32.8	31.6
e (climb one flight of stairs)	Normal	13.5	22.8	63.7
	High	18.3	27.5	54.3
g (walk more than a mile)	Normal	28.2	23.3	48.5
	High	35.1	24.5	40.4
h (walk half a mile)	Normal	21.4	16.1	62.5
	High	26.5	19.1	54.3
i (walk 100 yards)	Normal	10.9	15.3	73.8
	High	13.8	19.7	66.5

Table C8 Hypotheses 1g (low BMI): Results

Item	BMI group	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Normal	28.8	30.1	41.1
	Low	54.3	22.2	23.5
e (climb one flight of stairs)	Normal	13.5	22.8	63.7
	Low	31.6	29.0	39.4
g (walk more than a mile)	Normal	28.2	23.3	48.5
	Low	55.3	21.1	23.7
h (walk half a mile)	Normal	21.4	16.1	62.5
	Low	48.0	14.0	38.0
i (walk 100 yards)	Normal	10.9	15.3	73.8
	Low	24.7	26.0	49.3

Table C9 Hypotheses 2a (general health): Results

Item	General health	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Good	15.5	35.0	49.5
	Poor	66.1	25.2	8.7
e (climb one flight of stairs)	Good	4.9	18.5	76.6
	Poor	39.0	38.2	22.9
g (walk more than a mile)	Good	14.7	25.0	60.3
	Poor	66.5	21.8	11.7
h (walk half a mile)	Good	9.2	14.7	76.2
	Poor	53.9	24.1	22.0
i (walk 100 yards)	Good	3.3	10.1	86.7
	Poor	31.0	32.3	36.7

Table C10 Hypotheses 2b (participation restriction): Results

Item	Participation restriction	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Any	49.8	27.0	23.2
	None	15.0	36.6	48.4
e (climb one flight of stairs)	Any	28.4	32.9	38.7
	None	3.7	17.6	78.7
g (walk more than a mile)	Any	50.5	21.9	27.7
	None	13.4	26.5	60.0
h (walk half a mile)	Any	40.6	20.5	39.0
	None	7.5	15.1	77.4
i (walk 100 yards)	Any	22.8	25.9	51.3
	None	1.8	9.0	89.2

Table C11 Hypotheses 2c (aids or assistance inside the home): Results

Item	Aids or assistance	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Yes	87.1	9.4	3.5
	No	24.4	35.2	40.4
e (climb one flight of stairs)	Yes	66.0	28.4	5.6
	No	8.6	25.1	66.3
g (walk more than a mile)	Yes	90.0	6.3	3.7
	No	23.4	27.0	49.7
h (walk half a mile)	Yes	83.7	11.1	5.2
	No	15.0	19.2	65.8
i (walk 100 yards)	Yes	57.9	33.0	9.1
	No	5.4	15.4	79.2

Table C12 Hypotheses 2d (aids or assistance outside the home): Results

Item	Aids or assistance	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Yes	83.9	12.6	3.5
	No	21.3	36.0	42.7
e (climb one flight of stairs)	Yes	56.9	35.1	8.0
	No	7.3	23.3	69.4
g (walk more than a mile)	Yes	88.0	8.7	3.3
	No	19.8	27.7	52.6
h (walk half a mile)	Yes	79.1	14.7	6.3
	No	12.0	18.8	69.2
i (walk 100 yards)	Yes	48.5	37.9	13.7
	No	4.4	13.2	82.3

Table C13 Hypotheses 2e (live alone): Results

Item	Live alone	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Yes	45.3	29.9	24.8
	No	29.6	31.8	38.5
e (climb one flight of stairs)	Yes	25.1	31.1	43.9
	No	14.1	23.6	62.3
g (walk more than a mile)	Yes	45.2	23.3	31.5
	No	29.0	24.1	46.9
h (walk half a mile)	Yes	34.9	20.2	44.9
	No	21.7	17.1	61.2
i (walk 100 yards)	Yes	18.9	24.8	56.3
	No	11.1	15.5	73.4

Table C14 Hypotheses 3a (go out for a walk): Results

Item	Walk	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	17.4	32.2	50.4
	Less often	42.1	31.2	26.7
e (climb one flight of stairs)	Often	4.8	20.3	74.9
	Less often	23.5	28.3	48.2
g (walk more than a mile)	Often	12.2	24.2	63.7
	Less often	44.3	23.7	32.1
h (walk half a mile)	Often	6.2	15.2	78.7
	Less often	35.1	19.4	45.5
i (walk 100 yards)	Often	2.3	9.4	88.4
	Less often	18.9	22.6	58.4

Table C15 Hypotheses 3b (take bath/shower): Results

Item	Bath/shower	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	26.6	33.5	39.9
	Less often	50.5	26.6	22.9
e (climb one flight of stairs)	Often	11.3	24.3	64.4
	Less often	30.3	29.0	40.7
g (walk more than a mile)	Often	25.5	25.0	49.4
	Less often	51.6	21.1	27.4
h (walk half a mile)	Often	18.1	17.6	64.4
	Less often	41.8	18.9	39.3
i (walk 100 yards)	Often	8.3	15.7	76.0
	Less often	24.5	23.8	51.7

Table C16 Hypotheses 3c (heavy housework): Results

Item	Heavy housework	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	10.1	34.6	55.3
	Less often	35.2	31.3	33.5
e (climb one flight of stairs)	Often	4.3	16.9	78.8
	Less often	17.9	26.2	56.0
g (walk more than a mile)	Often	10.0	24.6	65.4
	Less often	34.8	23.8	41.4
h (walk half a mile)	Often	5.6	13.6	80.9
	Less often	26.5	18.2	55.3
i (walk 100 yards)	Often	2.9	6.3	90.8
	Less often	13.9	18.8	67.3

Table C17 Hypotheses 3d (heavy gardening): Results

Item	Heavy gardening	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	7.5	27.3	65.2
	Less often	34.5	31.6	33.9
e (climb one flight of stairs)	Often	2.7	14.4	82.9
	Less often	17.4	26.0	56.6
g (walk more than a mile)	Often	12.2	12.8	75.0
	Less often	34.0	24.1	41.9
h (walk half a mile)	Often	9.2	8.1	82.7
	Less often	25.8	18.3	56.0
i (walk 100 yards)	Often	4.3	5.4	90.3
	Less often	13.5	18.5	68.0

Table C18 Hypotheses 3e (DIY): Results

Item	DIY	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	3.9	27.3	68.8
	Less often	34.3	31.5	34.2
e (climb one flight of stairs)	Often	1.6	11.7	86.7
	Less often	17.3	25.9	56.8
g (walk more than a mile)	Often	7.0	13.3	79.7
	Less often	33.8	24.0	42.1
h (walk half a mile)	Often	1.6	10.2	88.3
	Less often	25.7	18.1	56.3
i (walk 100 yards)	Often	0.0	5.5	94.5
	Less often	13.4	18.3	68.2

Table C19 Hypotheses 3f (walk at least ¼ of a mile): Results

Item	Walk ¼ of a mile	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	14.6	32.3	53.1
	Less often	42.0	31.1	26.9
e (climb one flight of stairs)	Often	3.7	17.2	79.1
	Less often	22.9	29.3	47.8
g (walk more than a mile)	Often	8.6	22.2	69.2
	Less often	44.1	24.4	31.5
h (walk half a mile)	Often	3.6	12.7	83.7
	Less often	34.6	20.2	45.2
i (walk 100 yards)	Often	1.2	6.5	92.3
	Less often	18.4	23.0	58.5

Table C20 Hypotheses 3g (walk at least 2 miles): Results

Item	Walk 2 miles	% limited a lot	% limited a little	% not limited at all
d (climb several flights of stairs)	Often	9.1	26.5	64.4
	Less often	35.9	31.9	32.2
e (climb one flight of stairs)	Often	2.8	13.1	84.1
	Less often	18.2	26.8	55.0
g (walk more than a mile)	Often	3.5	14.2	82.3
	Less often	36.0	24.6	39.4
h (walk half a mile)	Often	1.7	8.0	90.4
	Less often	27.4	18.8	53.8
i (walk 100 yards)	Often	1.2	3.0	95.8
	Less often	14.3	19.4	66.3

Figure C1 Hypotheses 4 (number of participation domains restricted): Results

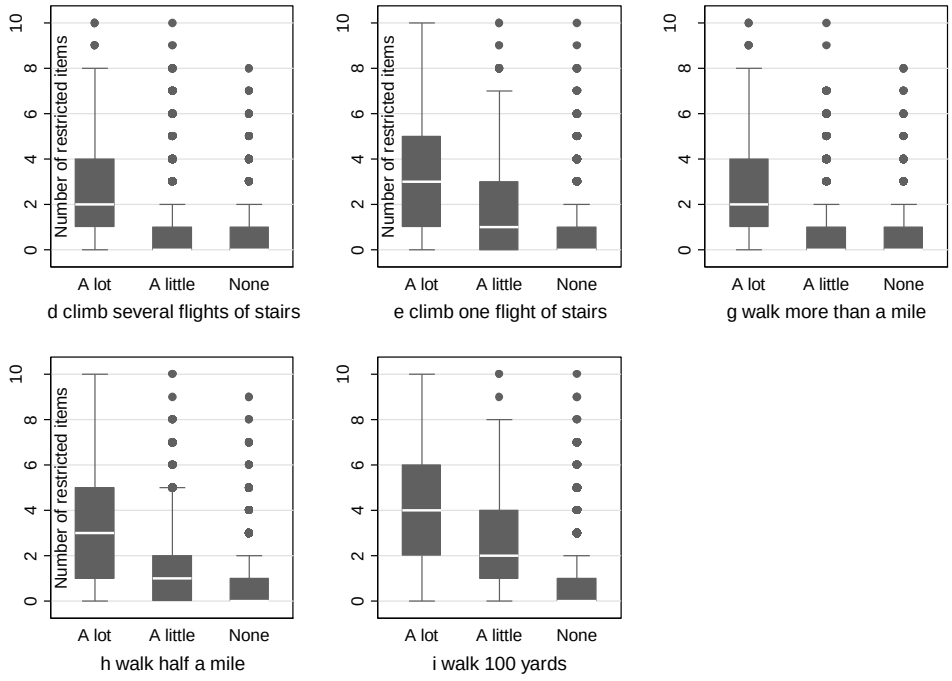


Figure C2 Hypotheses 5 (anxiety): Results

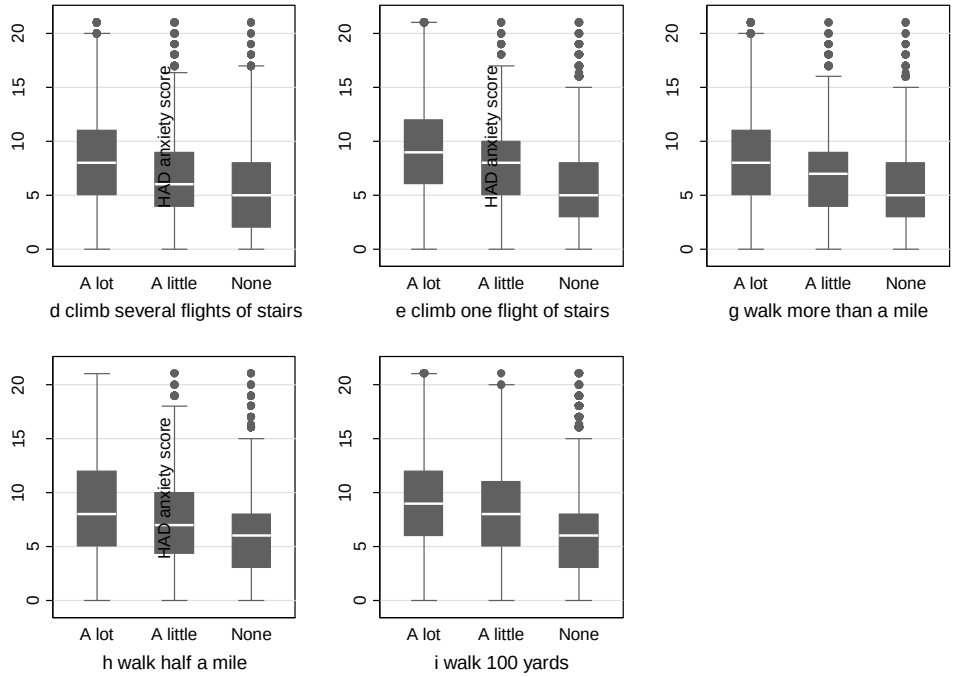


Figure C3 Hypotheses 6 (depression): Results

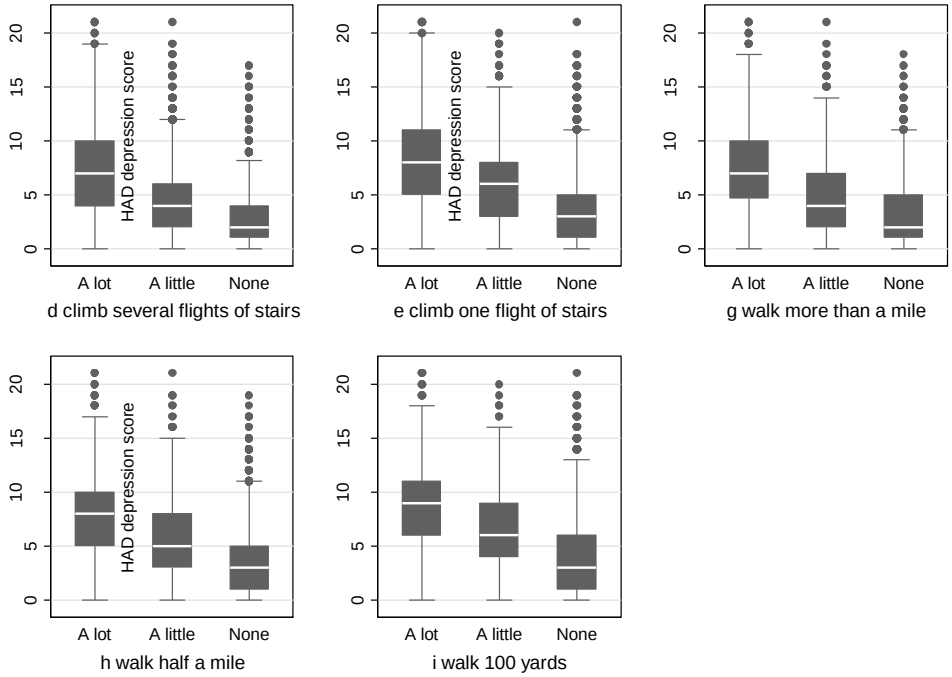


Figure C4 Hypotheses 7 (cognitive complaint): Results

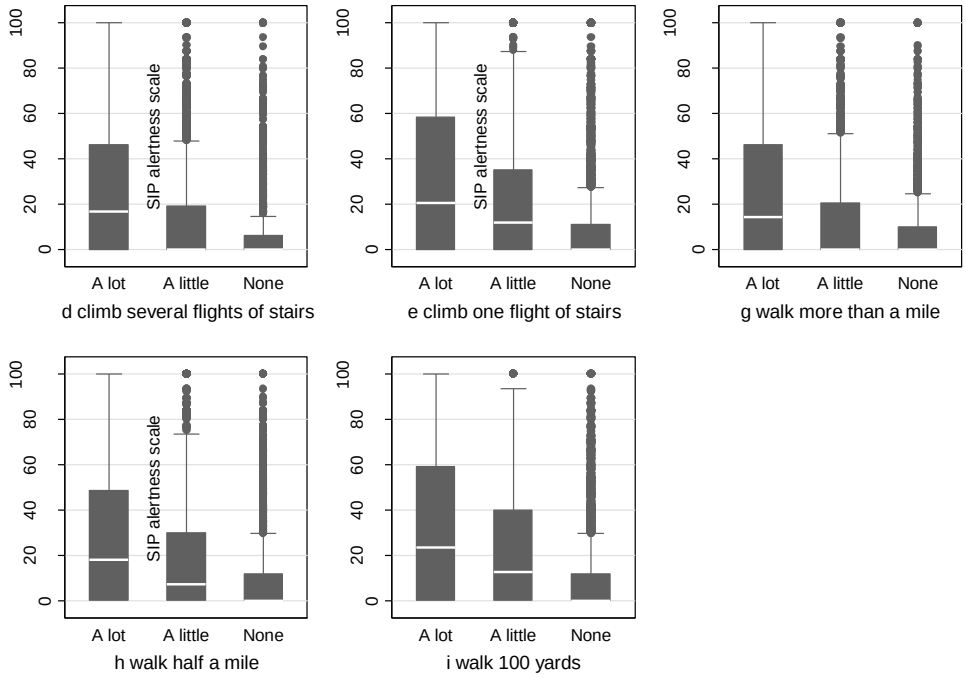
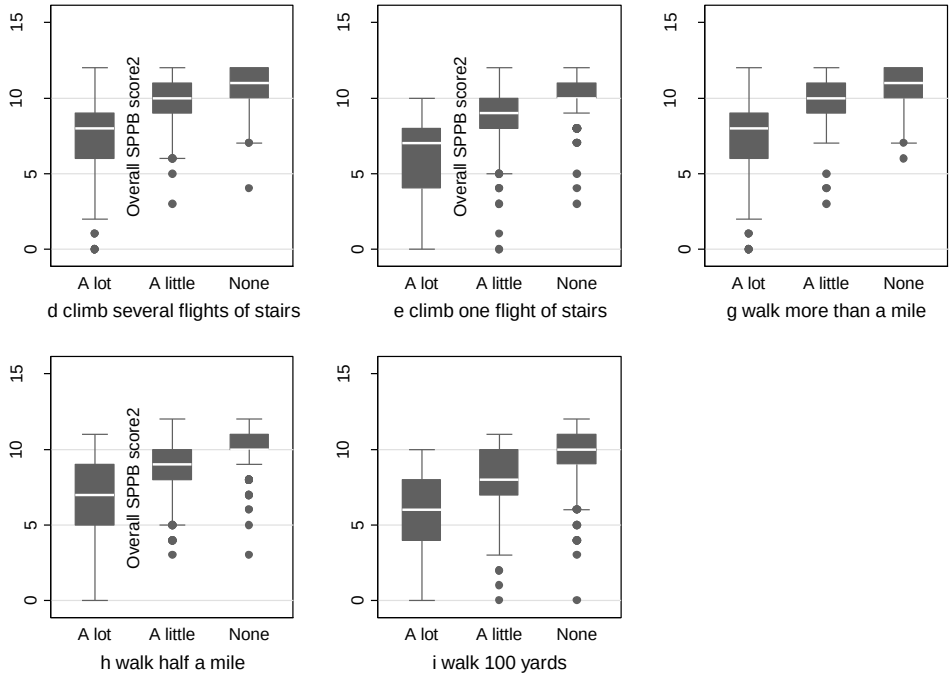


Figure C5 Hypotheses 8 (physical performance): Results



Appendix C4 Relative responses to locomotor disability-specific items within proposed hierarchies: walking and stair-climbing

Table C21 Relative responses to locomotor disability-specific items within proposed hierarchies: walking and stair climbing. NorStOP1 baseline cohort, n=7,878

		Climbing several flights of stairs		
		Limited a lot	Limited a little	Not limited
Climbing one flight of stairs	Limited a lot	1279 (16.7)	20 (0.3)	6 (0.1)
	Limited a little	1181 (15.5)	759 (9.9)	17 (0.2)
	Not limited	120 (1.6)	1614 (21.1)	2643 (34.6)
		Walking more than a mile		
		Limited a lot	Limited a little	Not limited
Walking half a mile	Limited a lot	1915 (25.2)	13 (0.2)	0 (0.0)
	Limited a little	582 (7.7)	760 (10.0)	18 (0.2)
	Not limited	39 (0.5)	1038 (13.6)	3245 (42.6)
		Walking half a mile		
		Limited a lot	Limited a little	Not limited
Walking 100 yards	Limited a lot	972 (12.8)	20 (0.3)	8 (0.1)
	Limited a little	831 (11.0)	507 (6.7)	32 (0.4)
	Not limited	112 (1.5)	824 (10.9)	4275 (56.4)

All values are n (%)

Appendix D Ordinal regression analysis of PF-10 locomotor disability-specific items

Appendix D1 Unstratified models of PF-10 locomotor disability-specific items

Table D1 Item d (climb several flights of stairs). OR (95%CI)

Independent variable	A little or a lot [a]	A lot [b]
Female	1.79 (1.65, 1.94)	1.49 (1.37, 1.63)
55 to 59 years	1.27 (1.12, 1.45)	1.41 (1.19, 1.67)
60 to 64 years	1.62 (1.41, 1.85)	2.04 (1.72, 2.41)
65 to 69 years	2.75 (2.39, 3.17)	2.55 (2.15, 3.03)
70 to 74 years	3.66 (3.15, 4.25)	4.06 (3.42, 4.83)
75 to 79 years	6.38 (5.35, 7.61)	5.85 (4.88, 7.01)
80 to 84 years	8.50 (6.75, 10.69)	9.25 (7.53, 11.37)
85 to 89 years	17.57 (11.54, 26.77)	13.78 (10.32, 18.41)
90 years and over	14.00 (7.08, 27.67)	36.52 (21.46, 62.17)
School only	1.30 (1.16, 1.45)	
Manual occupation	1.17 (1.07, 1.28)	1.31 (1.19, 1.44)
Inadequate income	1.86 (1.73, 1.99)	
Lives alone	1.17 (1.07, 1.27)	
Pain elsewhere only	1.62 (1.43, 1.83)	
Lower limb pain only	3.17 (2.79, 3.61)	
Lower limb and low back pain	2.77 (2.43, 3.15)	
Lower limb pain and pain elsewhere	4.55 (4.07, 5.08)	
Multiple pains	7.69 (6.96, 8.49)	

[a]Reference group is not limited; [b]Reference group is not limited or limited a little

Table D2 Item e (climb one flight of stairs). OR (95%CI)

Independent variable	A little or a lot [a]	A lot [b]
Female	1.49 (1.37, 1.62)	1.24 (1.11, 1.39)
55 to 59 years	1.33 (1.15, 1.55)	
60 to 64 years	1.86 (1.60, 2.16)	
65 to 69 years	2.58 (2.22, 3.00)	
70 to 74 years	3.98 (3.41, 6.64)	
75 to 79 years	6.15 (5.23, 7.24)	
80 to 84 years	10.03 (8.35, 12.05)	
85 to 89 years	16.85 (13.08, 21.71)	
90 years and over	42.79 (27.70, 66.12)	
School age education only	1.30 (1.14, 1.49)	
Manual occupation	1.46 (1.34, 1.60)	
Inadequate income	1.98 (1.83, 2.14)	
Lives alone	1.19 (1.08, 1.30)	
Pain elsewhere only	1.76 (1.50, 2.06)	
Lower limb pain only	3.77 (3.24, 4.39)	
Lower limb and low back pain	3.25 (2.79, 3.79)	
Lower limb pain and pain elsewhere	5.46 (4.80, 6.21)	
Multiple pains	8.62 (7.67, 9.68)	

[a]Reference group is not limited; [b]Reference group is not limited or limited a little

Table D3 Item g (walk more than a mile). OR (95%CI)

Independent variable	A little or a lot [a]	A lot [b]
Female	1.45 (1.34, 1.58)	1.20 (1.01, 1.31)
55 to 59 years	1.32 (1.16, 1.51)	1.23 (1.04, 1.46)
60 to 64 years	1.63 (1.42, 1.87)	1.78 (1.50, 2.10)
65 to 69 years	2.44 (2.12, 2.81)	2.30 (1.94, 2.72)
70 to 74 years	3.50 (3.01, 4.06)	3.98 (3.36, 4.73)
75 to 79 years	6.43 (5.43, 7.61)	6.58 (5.50, 7.88)
80 to 84 years	12.23 (9.76, 15.32)	11.09 (9.02, 13.63)
85 to 89 years	20.55 (14.12, 29.89)	21.70 (16.07, 29.30)
90 years and over	38.08 (18.01, 80.51)	95.42 (49.05, 185.64)
School age education only	1.28 (1.14, 1.44)	
Manual occupation	1.24 (1.14, 1.34)	
Inadequate income	1.90 (1.77, 2.05)	
Lives alone	1.13 (1.03, 1.23)	
Pain elsewhere only	1.67 (1.46, 1.90)	
Lower limb pain only	3.73 (3.26, 4.28)	
Lower limb and low back pain	3.49 (3.05, 4.00)	
Lower limb pain and pain elsewhere	5.03 (4.48, 5.64)	
Multiple pains	8.37 (7.55, 9.29)	

[a]Reference group is not limited; [b]Reference group is not limited or limited a little

Table D4 Item h (walk half a mile). OR (95%CI)

Independent variable	A little or a lot [a]	A lot [b]
Female	1.14 (1.06, 1.24)	
55 to 59 years	1.22 (1.06, 1.41)	
60 to 64 years	1.70 (1.47, 1.97)	
65 to 69 years	2.19 (1.89, 2.53)	
70 to 74 years	3.48 (3.00, 4.04)	
75 to 79 years	5.32 (4.54, 6.23)	
80 to 84 years	8.78 (7.30, 10.54)	
85 to 89 years	19.01 (14.54, 24.85)	
90 years and over	53.18 (32.05, 88.25)	
School age education only	1.24 (1.09, 1.42)	
Manual occupation	1.35 (1.24, 1.48)	
Inadequate income	2.01 (1.86, 2.17)	
Lives alone	1.12 (1.02, 1.23)	
Pain elsewhere only	1.85 (1.58, 2.16)	1.68 (1.37, 2.08)
Lower limb pain only	4.09 (3.50, 4.78)	3.66 (3.02, 4.43)
Lower limb and low back pain	3.59 (3.07, 4.20)	3.52 (2.91, 4.27)
Lower limb pain and pain elsewhere	5.45 (4.77, 6.23)	4.39 (3.72, 5.17)
Multiple pains	8.42 (7.47, 9.49)	7.61 (6.56, 8.81)

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Appendix D2 Socioeconomically stratified models: PF-10 Item i (walk 100 yards)

Table D5 Item i (walk 100 yards): stratified by educational attainment. OR (95%CI)

Independent variable	Further education		School-age education	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.17 (0.84, 1.63)		1.12 (1.02, 1.24)	0.96 (0.85, 1.10)
55 to 59 years	1.49 (0.91, 2.42)		1.20 (0.99, 1.44)	
60 to 64 years	1.70 (1.02, 2.85)		1.75 (1.45, 2.10)	
65 to 69 years	1.90 (1.05, 3.46)		2.16 (1.80, 2.60)	
70 to 74 years	3.94 (2.20, 7.06)		3.20 (2.66, 3.85)	
75 to 79 years	8.45 (4.62, 15.47)		4.56 (3.77, 5.52)	
80 to 84 years	17.89 (8.85, 36.16)		8.22 (6.67, 10.14)	
85 to 89 years	28.84 (12.52, 66.46)		13.68 (10.34, 18.09)	
90 years and over	49.82 (12.62, 196.72)		36.94 (23.98, 56.90)	
Manual occupation	1.86 (1.36, 2.55)		1.37 (1.24, 1.53)	
Inadequate income	2.69 (1.95, 3.70)		1.95 (1.78, 2.13)	
Lives alone	1.07 (0.75, 1.54)		1.21 (1.08, 1.35)	1.95 (1.78, 2.13)
Pain elsewhere only	1.32 (0.64, 2.71)		1.86 (1.54, 2.26)	
Lower limb pain only	3.02 (1.56, 5.84)		3.74 (3.12, 4.47)	
Lower limb and low back pain	6.46 (3.65, 11.45)		3.16 (2.62, 3.82)	
Lower limb pain and pain elsewhere	4.81 (2.82, 8.20)		4.57 (3.91, 5.33)	
Multiple pains	8.53 (5.36, 13.57)		7.98 (6.93, 9.18)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D6 Item i (walk 100 yards): stratified by occupational class. OR (95% CI)

Independent variable	Non-manual		Manual	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.14 (0.96, 1.36)		1.15 (0.97, 1.38)	1.08 (0.83, 1.39)
55 to 59 years	1.21 (0.86, 1.71)		1.21 (0.86, 1.71)	
60 to 64 years	1.86 (1.33, 2.62)		1.87 (1.33, 2.62)	
65 to 69 years	2.08 (1.47, 2.93)		2.08 (1.47, 2.93)	
70 to 74 years	3.30 (2.33, 4.66)		3.30 (2.34, 4.66)	
75 to 79 years	5.58 (3.92, 7.94)		5.58 (3.92, 7.95)	
80 to 84 years	11.65 (7.93, 17.10)		11.65 (7.94, 17.11)	
85 to 89 years	17.66 (10.70, 29.17)		17.67 (10.70, 29.18)	
90 years and over	44.18 (19.67, 99.21)		44.20 (19.70, 99.17)	
School age education only	1.45 (1.17, 1.81)		1.45 (1.17, 1.81)	
Inadequate income	2.22 (1.87, 2.64)		2.22 (1.87, 2.64)	
Lives alone	1.28 (1.04, 1.57)	0.85 (0.63, 1.13)	1.27 (1.03, 1.57)	0.86 (0.64, 1.16)
Pain elsewhere only	1.80 (1.23, 2.64)		1.80 (1.23, 2.64)	
Lower limb pain only	4.03 (2.88, 5.65)		4.03 (2.88, 5.65)	
Lower limb and low back pain	4.70 (3.37, 6.56)		4.70 (3.37, 6.56)	
Lower limb pain and pain elsewhere	5.19 (3.86, 6.98)		5.19 (3.86, 6.98)	
Multiple pains	8.65, 6.63, 11.30)		8.66 (6.63, 11.30)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D7 Item i (walk 100 yards): stratified by perceived adequacy of income. OR (95% CI)

Independent variable	Adequate		Inadequate	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.19 (1.04, 1.36)		1.08 (0.96, 1.23)	0.85 (0.72, 1.00)
55 to 59 years	1.22 (0.91, 1.64)		1.21 (0.97, 1.50)	1.52 (1.10, 2.10)
60 to 64 years	1.71 (1.27, 2.30)		1.82 (1.46, 2.27)	1.71 (1.23, 2.36)
65 to 69 years	2.77 (2.09, 3.67)		1.87 (1.49, 2.36)	1.85 (1.33, 2.58)
70 to 74 years	4.39 (3.32, 5.82)		2.66 (2.10, 3.38)	2.91 (2.09, 4.04)
75 to 79 years	7.41 (5.57, 9.85)		3.65 (2.84, 4.70)	2.97 (2.10, 4.19)
80 to 84 years	12.49 (9.18, 16.99)		7.09 (5.29, 9.50)	5.87 (4.11, 8.38)
85 to 89 years	21.55 (14.77, 31.45)		13.29 (8.24, 21.45)	8.29 (5.17, 13.30)
90 years and over	53.51 (31.90, 89.86)		22.96 (8.90, 59.20)	23.04 (10.40, 51.03)
School age education only	1.40 (1.12, 1.75)		1.12 (0.90, 1.40)	
Manual occupation	1.46 (1.27, 1.68)		1.35 (1.17, 1.56)	
Lives alone	1.18 (1.01, 1.38)	0.96 (0.77, 1.19)	1.14 (1.00, 1.30)	
Pain elsewhere only	1.79 (1.37, 2.33)		1.84 (1.42, 2.40)	
Lower limb pain only	4.02 (3.17, 5.10)		3.35 (2.61, 4.31)	
Lower limb and low back pain	3.52 (2.73, 4.54)		3.26 (2.54, 4.19)	
Lower limb pain and pain elsewhere	4.89 (3.94, 6.07)		4.39 (3.57, 5.40)	
Multiple pains	7.35 (6.04, 8.94)		8.45 (7.01, 10.19)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Appendix D3 **Stratified models: PF-10 items d (climb several flights of stairs), e (climb one flight of stairs), g (walk more than a mile) and h (walk half a mile)**

Appendix D3.1 Item d (climb several flights of stairs)

Table D8 Item d (climb several flights of stairs): stratified by gender. OR (95% CI)

Independent variable	Male		Female	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
55 to 59 years	1.37 (1.14, 1.65)		1.24 (1.04, 1.48)	1.35 (1.08, 1.67)
60 to 64 years	1.96 (1.63, 2.37)		1.49 (1.24, 1.80)	1.82 (1.46, 2.27)
65 to 69 years	2.77 (2.29, 3.35)		2.66 (2.19, 3.23)	2.49 (2.00, 3.10)
70 to 74 years	3.77 (3.08, 4.61)		3.63 (2.95, 4.46)	4.23 (3.39, 5.27)
75 to 79 years	5.07 (4.09. 6.29)		7.59 (5.86, 9.82)	6.95 (5.49, 8.81)
80 to 84 years	7.92 (6.05, 10.37)		9.70 (6.98, 13.47)	10.38 (7.94, 13.56)
85 to 89 years	11.59 (7.75, 17.32)		16.37 (9.35, 28.68)	17.95 (12.30, 26.20)
90 years and over	86.44 (23.75, 314.51)		10.95 (5.27, 22.75)	30.88 (17.17, 55.53)
School only	1.29 (1.08, 1.56)		1.30 (1.13, 1.50)	
Manual	1.35 (1.21, 1.52)		1.11 (1.00, 1.24)	
Inadequate	1.80 (1.59, 2.02)	2.09 (1.82, 2.39)	1.84 (1.68, 2.03)	
Lives alone	1.22 (1.06, 1.41)		1.11 (0.99, 1.24)	
Pain elsewhere only	1.63 (1.36, 1.96)		1.59 (1.35, 1.88)	
Lower limb pain only	2.52 (2.09, 3.05)		3.87 (3.24, 4.63)	
Lower limb and low back pain	2.20 (1.81, 2.66)		3.32 (2.78, 3.96)	
Lower limb pain and pain elsewhere	3.86 (3.27, 4.55)		5.20 (4.48, 6.03)	
Multiple pains	6.07 (5.23, 7.04)		9.27 (8.11, 10.60)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D9 (part I) Item d (climb several flights of stairs): stratified by age. OR (95% CI)

Independent variable	50 to 59 years		60 to 69 years	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.82 (0.59, 2.07)	1.42 (1.19, 1.68)	1.59 (1.39, 1.83)	1.26 (1.08, 1.47)
School age education only	1.41 (1.19, 1.67)		1.36 (1.12, 1.66)	
Manual occupation	1.15 (1.01, 1.31)		1.25 (1.09, 1.43)	
Inadequate income	2.00 (1.76, 2.29)	2.63 (2.21, 3.13)	1.85 (1.64, 2.10)	
Lives alone	1.29 (1.08, 1.53)		1.24 (1.06, 1.45)	
Pain elsewhere only	1.57 (1.26, 1.95)		1.52 (1.23,1.88)	
Lower limb pain only	3.26 (2.57, 4.15)		3.08 (2.46, 3.86)	
Lower limb and low back pain	2.73 (2.16, 3.47)		2.97 (2.38, 3.71)	
Lower limb pain and pain elsewhere	4.36 (3.57, 5.31)		4.84 (3.99, 5.86)	
Multiple pains	8.64 (7.28, 10.26)		7.73 (6.51, 9.19)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D9 (part II) Item d (climb several flights of stairs): stratified by age. OR (95% CI)

Independent variable	70 to 79 years		80 years and over	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.83 (1.58, 2.12)		1.93 (1.49, 2.89)	
School age education only	1.19 (0.91, 1.55)		1.04 (0.67, 1.61)	
Manual occupation	1.28 (1.10, 1.49)		1.26 (0.96, 1.66)	
Inadequate income	1.55 (1.35, 1.79)		1.93 (0.93, 2.84)	
Lives alone	1.09 (0.94, 1.27)		1.02 (0.80, 1.31)	
Pain elsewhere only	1.71 (1.33, 2.18)		1.63 (0.93, 2.84)	2.24 (1.43, 3.52)
Lower limb pain only	3.44 (2.67, 4.42)		3.48 (1.90, 6.37)	2.59 (1.73, 3.89)
Lower limb and low back pain	2.64 (2.05, 3.41)		1.94 (1.11, 3.40)	2.05 (1.36, 3.16)
Lower limb pain and pain elsewhere	4.23 (3.41, 5.25)		10.58 (4.79, 23.38)	3.95 (2.72, 5.74)
Multiple pains	6.53 (5.35, 7.96)		14.42 (6.18, 33.61)	4.80 (3.34, 6.90)

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D10 Item d (climb several flights of stairs): stratified by educational attainment. OR (95% CI)

Independent variable	Further education		School-age education	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.83 (1.48, 2.28)		1.78 (1.63, 1.95)	1.47 (1.34, 1.61)
55 to 59 years	1.51 (1.13, 2.03)		1.25 (1.08, 1.44)	1.36 (1.13, 1.63)
60 to 64 years	1.79 (1.30, 2.45)		2.66 (1.38, 1.85)	2.00 (1.67, 2.40)
65 to 69 years	3.16 (2.20, 4.53)		2.66 (2.29, 3.10)	2.47 (2.06, 2.96)
70 to 74 years	4.01 (2.70, 5.96)		3.62 (3.08, 4.25)	3.93 (3.27, 4.72)
75 to 79 years	9.38 (5.93, 14.84)		6.13 (5.08, 7.40)	5.48 (4.52, 6.64)
80 to 84 years	9.12 (5.01, 16.62)		8.22 (6.45, 10.46)	9.08 (7.31, 11.29)
85 to 89 years	23.59 (11.13, 50.00)		17.46 (10.98, 27.76)	12.52 (9.20, 17.03)
90 years and over	136.02 (15.12, 1223.56)		12.05 (6.05, 23.99)	31.68 (18.26, 54.93)
Manual occupation	0.99 (0.78, 1.26)	1.42 (1.06, 1.90)	1.24 (1.14, 1.35)	
Inadequate income	2.20 (1.76, 2.75)		1.82 (1.69, 1.97)	
Lives alone	1.20 (0.92, 1.55)		1.16 (1.06, 1.28)	
Pain elsewhere only	1.26 (0.88, 1.81)		1.69 (1.48, 1.92)	
Lower limb pain only	2.90 (1.94, 4.32)		3.22 (2.81, 3.69)	
Lower limb and low back pain	4.00 (2.76, 5.79)		2.64 (2.30, 3.03)	
Lower limb pain and pain elsewhere	3.99 (2.88, 5.52)		4.64 (4.13, 5.22)	
Multiple pains	7.12 (5.36, 9.47)		7.77 (6.99, 8.64)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D11 Item d (climb several flights of stairs): stratified by occupational class. OR (95% CI)

Independent variable	Non-manual		Manual	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.95 (1.72, 2.22)		1.66 (1.51, 1.84)	1.38 (1.25, 1.53)
55 to 59 years	1.42 (1.14, 1.77)	1.75 (1.26, 2.43)	1.21 (1.03, 1.42)	1.30 (1.06, 1.58)
60 to 64 years	1.63 (1.29, 2.04)	2.67 (1.93, 3.71)	1.63 (1.38, 1.92)	1.86 (1.53, 2.27)
65 to 69 years	2.90 (2.28, 3.68)	2.94 (2.11, 4.09)	2.69 (2.26, 3.20)	2.45 (2.00, 2.99)
70 to 74 years	3.59 (2.79, 4.63)	4.71 (3.37, 6.57)	3.73 (3.10, 4.50)	3.91 (3.19, 4.79)
75 to 79 years	8.18 (6.04, 11.08)	7.64 (5.39, 10.83)	5.54 (4.55, 7.00)	5.32 (4.30, 6.59)
80 to 84 years	10.32 (6.86, 15.53)	11.15 (7.49, 16.60)	7.85 (5.94, 10.36)	8.81 (6.91, 11.24)
85 to 89 years	22.62 (11.23, 45.57)	18.66 (11.03, 31.55)	15.35 (9.08, 25.94)	12.54 (8.85, 17.76)
90 years and over	25.37 (5.74, 112.24)	44.01 (16.62, 116.59)	11.62 (5.39, 25.07)	34.60 (18.28, 65.48)
School education only	1.27 (1.10, 1.47)		1.35 (1.13, 1.61)	
Inadequate income	2.05 (1.80, 2.34)		1.79 (1.64, 1.94)	
Lives alone	1.18 (1.01, 1.38)		1.16 (1.04, 1.28)	
Pain elsewhere only	1.73 (1.39, 2.15)		1.58 (1.36, 1.83)	
Lower limb pain only	3.32 (2.64, 4.17)		3.12 (2.67, 3.66)	
Lower limb and low back pain	3.19 (2.55, 3.98)		2.58 (2.20, 3.02)	
Lower limb pain and pain elsewhere	4.44 (3.66, 5.40)		4.61 (4.03, 5.27)	
Multiple pains	7.90 (6.64, 9.41)		7.59 (6.72, 8.56)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D12 Item d (climb several flights of stairs): stratified by perceived adequacy of income. OR (95% CI)

Independent variable	Adequate		Inadequate	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.75 (1.59, 1.93)		1.75 (1.54, 2.00)	1.38 (1.22, 1.56)
55 to 59 years	1.33 (1.12, 1.57)		1.21 (0.99, 1.47)	1.44 (1.15, 1.78)
60 to 64 years	1.72 (1.44, 2.04)		1.62 (1.32, 1.99)	2.01 (1.61, 2.51)
65 to 69 years	3.06 (2.58, 3.63)		2.40 (1.91, 3.00)	2.08 (1.65, 2.61)
70 to 74 years	4.37 (3.65, 5.22)		3.11 (2.46, 3.94)	3.56 (2.81, 4.49)
75 to 79 years	7.51 (6.18, 9.13)		5.31 (4.00, 7.06)	4.33 (3.38, 5.56)
80 to 84 years	10.73 (8.49, 13.56)		6.42 (4.45, 9.27)	7.50 (5.59, 10.07)
85 to 89 years	19.12 (13.70, 26.69)		17.00 (7.24, 39.90)	8.13 (5.18, 12.77)
90 years and over	41.95 (22.47, 78.30)		7.10 (2.06, 24.52)	20.17 (7.69, 52.37)
School age education only	1.31 (1.14, 1.51)		1.26 (1.05, 1.51)	
Manual occupation	1.27 (1.15, 1.40)		1.15 (1.02, 1.30)	
Lives alone	1.08 (0.95, 1.22)		1.24 (1.10, 1.40)	
Pain elsewhere only	1.50 (1.27, 1.76)		1.82 (1.50, 2.21)	
Lower limb pain only	2.93 (2.78, 3.46)		3.58 (2.91, 4.40)	
Lower limb and low back pain	2.44 (2.05, 2.89)		3.32 (2.72, 4.05)	
Lower limb pain and pain elsewhere	4.21 (3.63, 4.88)		5.13 (4.34, 6.07)	
Multiple pains	6.69 (5.86, 7.65)		9.02 (7.76, 10.48)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D13 Item e (climb one flight of stairs): stratified by gender. OR (95% CI)

Independent variable	Male		Female	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
55 to 59 years	1.40 (1.10, 1.77)		1.29 (1.06, 1.56)	
60 to 64 years	2.11 (1.67, 2.67)		1.69 (1.39, 2.06)	
65 to 69 years	2.54 (2.01, 3.23)		2.63 (2.16, 3.20)	
70 to 74 years	3.64 (2.84, 4.66)		4.29 (3.51, 5.23)	
75 to 79 years	5.44 (4.21, 7.02)		6.88 (5.57, 8.51)	
80 to 84 years	8.62 (6.43, 11.56)		11.45 (9.03, 14.51)	
85 to 89 years	16.39 (10.83, 24.81)		18.01 (13.04, 24.87)	
90 years and over	80.94 (30.93, 211.85)		38.45 (23.59, 62.67)	
School age education only	1.58 (1.23, 2.02)		1.19 (1.01, 1.41)	
Manual occupation	1.51 (1.32, 1.73)		1.35 (1.19, 1.53)	1.61 (1.35, 1.91)
Inadequate income	2.18 (1.93, 2.47)		1.87 (1.68, 2.07)	
Lives alone	1.40 (1.18, 1.65)	1.09 (0.87, 1.37)	1.10 (0.98, 1.24)	
Pain elsewhere only	1.62 (1.27, 2.06)		1.88 (1.53, 2.31)	
Lower limb pain only	3.01 (2.39, 3.79)		4.52 (3.69, 5.54)	
Lower limb and low back pain	2.52 (1.99, 3.20)		3.94 (3.22, 4.83)	
Lower limb pain and pain elsewhere	4.55 (3.73, 5.55)		6.24 (5.27, 7.40)	
Multiple pains	6.80 (5.70, 8.13)		10.25 (8.78, 11.96)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D14 (part I) Item e (climb one flight of stairs): stratified by age group. OR (95% CI)

Independent variable	50 to 59 years		60 to 69 years	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.45 (1.24, 1.69)	1.11 (0.86, 1.42)	1.32 (1.15, 1.51)	
School age education only	1.32 (1.06, 1.65)		1.46 (1.14, 1.87)	
Manual	1.35 (1.14, 1.60)		1.50 (1.28, 1.75)	
Inadequate	2.46 (2.10, 2.88)	3.52 (2.64, 4.69)	2.11 (1.83, 2.42)	
Lives alone	1.44 (1.17, 1.76)	1.09 (0.92, 1.30)	1.18 (1.01, 1.37)	
Pain elsewhere only	1.66 (1.16, 2.37)	3.30 (1.58, 6.90)	1.45 (1.08, 1.95)	
Lower limb pain only	3.90 (2.75, 5.52)	6.21 (3.04, 12.70)	3.73 (2.82, 4.94)	
Lower limb and low back pain	4.01 (2.86, 5.64)	4.02 (1.87, 8.68)	3.49 (2.66, 4.57)	
Lower limb pain and pain elsewhere	6.60 (4.96, 8.78)	6.38 (3.32, 12.23)	5.75 (4.54, 7.27)	
Multiple pains	10.86 (8.40, 14.02)	14.67 (8.12, 26.48)	8.70 (7.04, 10.76)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D14 (part II) Item e (climb one flight of stairs): stratified by age group. OR (95% CI)

Independent variable	70 to 79 years		80 years and over	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.68 (1.43, 1.98)	1.34 (1.10, 1.63)	1.64 (1.29, 2.08)	
School age education only	1.23 (0.92, 1.65)		1.06 (0.70, 1.62)	
Manual	1.57 (1.33, 1.85)		1.36 (1.05, 1.75)	
Inadequate	1.55 (1.34, 1.79)		1.47 (1.17, 1.84)	
Lives alone	1.10 (0.88, 1.38)		1.10 (0.88, 1.38)	
Pain elsewhere only	1.91 (1.46, 2.51)		2.53 (1.67, 3.82)	
Lower limb pain only	3.73 (2.87, 4.85)		4.05 (2.78, 5.89)	
Lower limb and low back pain	2.72 (2.06, 3.59)		2.64 (1.78, 3.92)	
Lower limb pain and pain elsewhere	4.59 (3.66, 5.74)		5.04 (3.62, 7.03)	
Multiple pains	6.75 (5.50, 8.29)		7.21 (5.22, 9.96)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D15 Item e (climb one flight of stairs): stratified by educational attainment. OR (95% CI)

Independent variable	Further education		School-age education	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	2.01 (1.50, 2.68)		1.45 (1.33, 1.58)	1.21 (1.07, 1.36)
55 to 59 years	1.61 (1.08, 2.40)		1.29 (1.10, 1.52)	
60 to 64 years	1.83 (1.19, 2.80)		1.85 (1.57, 2.17)	
65 to 69 years	2.71 (1.69, 4.35)		2.53 (2.16, 2.97)	
70 to 74 years	4.41 (2.67, 7.29)		3.89 (3.30, 4.58)	
75 to 79 years	9.66 (5.66, 16.50)		5.86 (4.94, 6.96)	
80 to 84 years	12.51 (6.38, 24.52)		9.71 (8.01, 11.77)	
85 to 89 years	37.36 (17.38, 80.33)		15.39 (11.75, 20.14)	
90 years and over	49.88 (13.23, 188.03)		41.83 (26.32, 66.47)	
Lives alone	1.06 (0.77, 1.46)		1.20 (1.09, 1.32)	
Manual	1.41 (1.08, 1.86)		1.47 (1.34, 1.62)	
Inadequate	2.20 (1.66, 2.94)	3.29 (2.14, 5.03)	1.94 (1.79, 2.11)	
Pain elsewhere only	1.20 (0.65, 2.23)		1.81 (1.54, 2.14)	
Lower limb pain only	4.76 (2.78, 8.14)		3.72 (3.17, 4.35)	
Lower limb and low back pain	6.91 (4.20, 11.36)		3.01 (2.56, 3.54)	
Lower limb pain and pain elsewhere	6.45 (4.13, 10.08)		5.39 (4.71, 6.17)	
Multiple pains	12.58 (8.46, 18.71)		8.29 (7.34, 9.36)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D16 Item e (climb one flight of stairs): stratified by occupational class. OR (95% CI)

Independent variable	Non-manual		Manual	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.76 (1.50, 2.06)	1.26 (1.00, 1.60)	1.35 (1.23, 1.48)	
55 to 59 years	1.43 (1.08, 1.89)		1.30 (1.09, 1.55)	
60 to 64 years	1.87 (1.40, 2.49)		1.87 (1.56, 2.23)	
65 to 69 years	2.62 (1.97, 3.49)		2.57 (2.15, 3.07)	
70 to 74 years	3.83 (2.85, 5.13)		4.05 (3.37, 4.85)	
75 to 79 years	7.21 (5.31, 9.79)		5.79 (4.78, 7.02)	
80 to 84 years	11.30 (7.97, 16.02)		9.62 (7.75, 11.94)	
85 to 89 years	27.76 (17.53, 43.96)		13.56 (10.03, 18.35)	
90 years and over	47.94 (21.96, 104.66)		41.03 (24.21, 69.54)	
Lives alone	1.07 (0.89, 1.28)		1.23 (1.10, 1.36)	
School only	1.31 (1.09, 1.57)		1.30 (1.06, 1.60)	
Inadequate	2.19 (1.88, 2.55)		1.90 (1.74, 2.09)	
Pain elsewhere only	1.84 (1.35, 2.51)		1.74 (1.44, 2.09)	
Lower limb pain only	4.20 (3.15, 5.59)		3.64 (3.05, 4.35)	
Lower limb and low back pain	3.81 (2.85, 5.08)		3.04 (2.54, 3.65)	
Lower limb pain and pain elsewhere	6.08 (4.75, 7.78)		5.25 (4.52, 6.11)	
Multiple pains	9.81 (7.85, 12.25)		8.18 (7.13, 9.38)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D17 Item e (climb one flight of stairs): stratified by perceived adequacy of income. OR (95% CI)

Independent variable	Adequate		Inadequate	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.68 (1.49, 1.89)	1.27 (1.07, 1.52)	1.59 (1.42, 1.79)	
55 to 59 years	1.25 (0.99, 1.58)		1.25 (0.99, 1.58)	
60 to 64 years	1.84 (1.46, 2.33)		1.84 (1.45, 2.32)	
65 to 69 years	3.05 (2.43, 3.82)		3.04 (2.42,3.80)	
70 to 74 years	4.97 (3.95, 6.25)		4.94 (3.92, 6.21)	
75 to 79 years	8.13 (6.40, 10.33)		8.09 (6.36, 10.28)	
80 to 84 years	13.05 (10.00, 17.04)		12.93 (9.91, 16.88)	
85 to 89 years	22.22 (15.79, 31.28)		22.06 (15.67, 31.05)	
90 years and over	54.79 (32.39, 92.69)		53.56 (31.61, 90.76)	
Lives alone	1.15 (1.00, 1.32)		1.14 (0.99, 1.31)	
School only	1.34 (1.12, 1.61)		1.34 (1.12, 1.61)	
Manual	1.53 (1.36, 1.72)		1.52 (1.35, 1.72)	
Pain elsewhere only	1.68 (1.35, 2.09)		1.67 (1.34, 2.08)	
Lower limb pain only	3.75 (3.05, 4.61)		3.73 (3.03, 4.56)	
Lower limb and low back pain	3.21 (2.58, 3.98)		3.19 (2.57, 3.96)	
Lower limb pain and pain elsewhere	5.62 (4.69, 6.74)		5.59 (4.66, 6.70)	
Multiple pains	8.45 (7.17, 9.96)		8.39 (7.12, 9.89)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D18 Item g (walk more than a mile): stratified by gender. OR (95% CI)

Independent variable	Male		Female	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
55 to 59 years	1.29 (1.06, 1.56)		1.37 (1.12, 1.67)	1.08 (0.83, 1.41)
60 to 64 years	1.81 (1.49, 2.21)		1.75 (1.42, 2.15)	1.91 (1.48, 2.46)
65 to 69 years	2.32 (1.90, 2.83)		2.39 (1.94, 2.95)	2.17 (1.68, 2.81)
70 to 74 years	3.47 (2.81, 4.28)		3.34 (2.67, 4.18)	3.57 (2.75, 4.65)
75 to 79 years	5.40 (4.31, 6.77)		5.29 (4.14, 6.76)	5.38 (4.10, 7.08)
80 to 84 years	8.94 (6.77, 11.80)		9.52 (6.87, 13.21)	8.46 (6.14, 11.65)
85 to 89 years	15.74 (10.20, 24.29)		16.93 (9.71, 29.51)	15.10 (9.42, 24.20)
90 years and over	89.73 (24.89, 323.45)		94.40 (12.25, 727.77)	87.85 (24.03, 321.20)
Lives alone	1.22 (1.05, 1.42)		1.22 (1.05, 1.41)	
School only	1.28 (1.05, 1.55)		1.28 (1.05, 1.56)	
Manual	1.30 (1.15, 1.46)		1.25 (1.10, 1.42)	1.39 (1.20, 1.61)
Inadequate	2.05 (1.83, 2.29)		2.05 (1.83, 2.29)	
Pain elsewhere only	1.77 (1.45, 2.15)		1.76 (1.44, 2.15)	
Lower limb pain only	3.60 (2.95, 4.38)		3.58 (2.94, 4.37)	
Lower limb and low back pain	3.23 (2.64, 3.94)		3.21 (2.63, 3.92)	
Lower limb pain and pain elsewhere	4.80 (4.03, 5.72)		4.78 (4.02, 5.70)	
Multiple pains	7.15 (6.11, 8.40)		7.13 (6.10, 8.34)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D19 Item g (walk more than a mile): stratified by age group. OR (95% CI)

Independent variable	50 to 59 years		60 to 69 years	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.31 (1.15, 1.49)		1.36 (1.18, 1.55)	1.03 (0.88, 1.20)
Lives alone	1.31 (1.09, 1.68)		1.04 (0.89, 1.22)	
School only	1.40 (1.17, 1.68)		1.35 (1.09, 1.67)	
Manual	1.19 (1.03, 1.37)		1.35 (1.17, 1.55)	
Inadequate	2.09 (1.83, 2.40)	2.75 (2.30, 3.28)	2.07 (1.82, 2.35)	
Pain elsewhere only	1.92 (1.50, 2.47)		1.41 (1.12, 1.79)	
Lower limb pain only	4.24 (3.26, 5.51)		3.72 (2.94, 4.71)	
Lower limb and low back pain	3.58 (2.75, 4.64)		3.60 (2.86, 4.54)	
Lower limb pain and pain elsewhere	5.70 (4.58, 7.09)		5.46 (4.47, 6.68)	
Multiple pains	11.19 (9.23, 13.57)		8.34 (6.95, 10.00)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D19 cont. Item g (walk more than a mile): stratified by age group. OR (95% CI)

Independent variable	70 to 69 years		80 years and over	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.62 (1.37, 1.91)	1.30 (1.11, 1.53)	1.87 (1.44, 2.43)	
Lives alone	1.18 (1.01, 1.37)		0.78 (0.55, 1.10)	1.06 (0.81, 1.38)
School only	1.12 (0.86, 1.45)		1.11 (0.70, 1.77)	
Manual	1.22 (1.04, 1.42)		1.06 (0.80, 1.41)	
Inadequate	1.46 (1.26, 1.68)		1.27 (0.99, 1.64)	
Pain elsewhere only	2.00 (1.55, 2.56)		1.30 (0.86, 1.98)	
Lower limb pain only	3.87 (3.01, 4.97)		2.97 (1.98, 4.47)	
Lower limb and low back pain	3.51 (2.71, 4.55)		2.64 (1.72, 4.07)	
Lower limb pain and pain elsewhere	4.45 (3.58, 5.52)		3.40 (2.37, 4.88)	
Multiple pains	6.10 (5.01, 7.42)		5.17 (3.56, 7.51)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D20 Item g (walk more than a mile): stratified by educational attainment. OR (95% CI)

Independent variable	Further education*		School-age education	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.39 (0.99, 1.94)		1.45 (1.34, 1.58)	1.18 (1.08, 1.30)
55 to 59 years	1.39 (1.00, 1.94)	1.45 (0.92, 2.30)	1.31 (1.13, 1.52)	1.20 (1.00, 1.44)
60 to 64 years	1.75 (1.22, 2.50)	2.01 (1.25, 3.24)	1.60 (1.38, 1.86)	1.74 (1.45, 2.08)
65 to 69 years	2.44 (1.61, 3.71)	2.40 (1.41, 4.08)	2.41 (2.07, 2.81)	2.25 (1.88, 2.70)
70 to 74 years	4.13 (2.62, 6.50)	5.38 (3.15, 9.20)	3.41 (2.91, 4.00)	3.84 (3.20, 4.61)
75 to 79 years	15.73 (9.02, 27.41)	9.76 (5.36, 17.77)	5.90 (4.93, 7.05)	6.27 (5.18, 7.59)
80 to 84 years	13.78 (6.36, 29.86)	15.72 (7.59, 32.55)	11.93 (9.40, 15.13)	10.64 (8.56, 13.22)
85 to 89 years	39.00 (13.80, 110.19)	49.23 (20.74, 116.90)	18.79 (12.57, 28.10)	19.45 (14.13, 26.77)
90 years and over	‡	‡	44.31 (18.89, 103.92)	99.44 (48.04, 205.85)
Lives alone	0.97 (0.74, 1.28)		1.15 (1.05, 1.26)	
Manual	1.48 (1.17, 1.88)		1.25 (1.15, 1.37)	
Inadequate	1.14 (0.90, 1.45)		1.86 (1.72, 2.00)	
Pain elsewhere only	1.22 (0.80, 1.88)		1.73 (1.51, 2.00)	
Lower limb pain only	5.14 (3.35, 7.87)		3.64 (3.15, 4.19)	
Lower limb and low back pain	5.12 (3.41, 7.70)		3.35 (2.90, 3.86)	
Lower limb pain and pain elsewhere	4.82 (3.38, 6.88)		5.09 (4.51, 5.75)	
Multiple pains	9.89 (7.22, 13.56)		8.26 (7.40, 9.22)	

*This model predicts 10 negative probabilities; ‡Inestimable; [a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D21 Item g (walk more than a mile): stratified by occupational class. OR (95% CI)

Independent variable	Non-manual		Manual	
	A little or a lot [a]	A lot [b]	A little or a lot [a]	A lot [b]
Female	1.54 (1.35, 1.76)		1.43 (1.30, 1.58)	1.10 (0.99, 1.23)
55 to 59 years	1.48 (1.18, 1.87)	1.33 (0.96, 1.83)	1.23 (1.06, 1.43)	
60 to 64 years	1.59 (1.25, 2.02)	2.01 (1.46, 2.76)	1.68 (1.43, 1.96)	
65 to 69 years	2.48 (1.94, 3.17)	2.41 (1.75, 3.33)	2.36 (1.02, 2.77)	
70 to 74 years	3.59 (2.77, 4.65)	5.03 (3.65, 6.94)	3.59 (3.05, 4.24)	
75 to 79 years	8.92 (6.62, 12.03)	8.16 (5.81, 11.45)	5.84 (4.88, 7.00)	
80 to 84 years	14.69 (9.85, 21.90)	15.12 (10.23, 22.33)	10.33 (8.33, 12.81)	
85 to 89 years	29.92 (15.72, 56.97)	34.30 (20.00, 58.85)	17.92 (12.77, 25.15)	
90 years and over	33.39 (9.31, 119.66)	98.02 (30.61, 313.91)	91.82 (40.63, 207.52)	
Lives alone	1.00 (0.84, 1.18)		1.19 (1.07, 1.32)	
School only	1.24 (1.06, 1.45)		1.33 (1.11, 1.60)	
Inadequate	2.13 (1.85, 2.44)		1.82 (1.67, 1.99)	
Pain elsewhere only	1.70 (1.34, 2.16)		1.66 (1.41, 1.94)	
Lower limb pain only	4.44 (3.49, 5.65)		3.48 (2.95, 4.09)	
Lower limb and low back pain	4.29 (3.38, 5.44)		3.18 (2.70, 3.75)	
Lower limb pain and pain elsewhere	5.11 (4.15, 6.29)		5.02 (4.67, 5.76)	
Multiple pains	8.54 (7.10, 10.28)		8.31 (7.33, 9.42)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D22 Item g (walk more than a mile): stratified by perceived adequacy of income. OR (95% CI)

Independent variable	Adequate		Inadequate	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
Female	1.57 (1.41, 1.75)	1.31 (1.15, 1.49)	1.32 (1.17, 1.50)	1.10 (0.97, 1.25)
55 to 59 years	1.31 (1.09, 1.58)	1.17 (0.89, 1.56)	1.31 (1.09, 1.56)	
60 to 64 years	1.46 (1.21, 1.78)	1.91 (1.45, 2.51)	1.81 (1.51, 2.17)	
65 to 69 years	2.73 (2.26, 3.30)	2.76 (2.11, 3.60)	2.07 (1.71, 2.51)	
70 to 74 years	4.02 (3.29, 4.91)	5.15 (3.96, 6.71)	3.19 (2.61, 3.89)	
75 to 79 years	8.26 (6.59, 10.35)	9.30 (7.08, 12.22)	4.80 (3.85, 5.97)	
80 to 84 years	14.05 (10.50, 18.79)	15.60 (11.53, 21.11)	8.61 (6.57, 11.27)	
85 to 89 years	22.59 (14.47, 35.27)	29.89 (20.11, 44.42)	15.74 (9.87, 25.10)	
90 years and over	52.81 (20.69, 134.78)	154.38 (66.92, 356.15)	35.08 (11.63, 105.75)	
Lives alone	1.10 (0.97, 1.25)		1.14 (1.01, 1.29)	
School only	1.34 (1.15, 1.56)		1.18 (0.98, 1.42)	
Manual	1.28 (1.15, 1.42)		1.18 (1.04, 1.33)	
Pain elsewhere only	1.53 (1.28, 1.83)		1.85 (1.51, 2.26)	
Lower limb pain only	3.72 (3.11, 4.44)		3.79 (3.08, 4.67)	
Lower limb and low back pain	3.16 (2.64, 3.79)		3.96 (3.23, 4.86)	
Lower limb pain and pain elsewhere	4.83 (4.13, 4.66)		5.34 (4.51, 6.33)	
Multiple pains	7.52 (6.64, 8.66)		9.39 (8.05, 10.95)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D23 Item h (walking half a mile): stratified by gender. OR (95% CI)

Independent variable	Male		Female	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
55 to 59 years	1.23 (0.99, 1.53)		1.21 (1.00, 1.46)	
60 to 64 years	1.81 (1.45, 2.25)		1.62 (1.33, 1.97)	
65 to 69 years	2.23 (1.79, 2.78)		2.17 (1.79, 2.63)	
70 to 74 years	3.36 (2.67, 4.23)		3.64 (2.99, 4.44)	
75 to 79 years	4.82 (3.78, 6.13)		5.93 (4.80, 7.33)	
80 to 84 years	6.41 (4.81, 8.54)		11.27 (8.85, 14.35)	
85 to 89 years	12.03 (7.89, 18.35)		27.46 (19.23, 39.19)	
90 years and over	59.88 (20.29, 176.74)		56.88 (31.85, 101.59)	
Lives alone	1.29 (1.10, 1.50)		1.00 (0.89, 1.13)	
School only	1.21 (0.97, 1.51)		1.26 (1.07, 1.48)	
Manual	1.39 (1.22, 2.59)		1.29 (1.15, 1.45)	
Inadequate	2.18 (1.94, 2.46)		1.91 (1.72, 2.12)	
Pain elsewhere only	1.94 (1.55, 2.44)	1.67 (1.23, 2.27)	1.75 (1.42, 2.16)	
Lower limb pain only	3.62 (2.89, 4.53)	2.73 (2.05, 3.63)	4.66 (3.80, 5.73)	
Lower limb and low back pain	2.98 (2.36, 3.75)	2.79 (2.09, 3.73)	4.29 (3.50, 5.26)	
Lower limb pain and pain elsewhere	5.09 (4.17, 6.21)	3.55 (2.77, 4.55)	5.63 (4.74, 6.68)	
Multiple pains	7.02 (5.87, 8.39)	6.50 (5.22, 8.08)	9.43 (8.07, 11.02)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D24 Item h (walking half a mile): stratified by age group. OR (95% CI)

Independent variable	50 to 59 years		60 to 69 years	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
Female	1.07 (0.92, 1.24)		1.02 (0.89, 1.17)	
Lives alone	1.44 (1.18, 1.75)		1.07 (0.90, 1.27)	
School only	1.33 (1.08, 1.65)		1.30 (1.02, 1.65)	
Manual	1.37 (1.16, 1.62)		1.41 (1.21, 1.65)	
Inadequate	2.43 (2.09, 2.83)	3.06 (2.47, 3.79)	2.20 (1.92, 2.53)	
Pain elsewhere only	2.28 (1.65, 3.14)		1.46 (1.09, 1.86)	
Lower limb pain only	4.36 (3.13, 6.06)		3.79 (2.88, 4.99)	
Lower limb and low back pain	4.04 (2.91, 5.61)		3.45 (2.63, 4.51)	
Lower limb pain and pain elsewhere	6.19 (4.70, 8.16)		5.40 (4.28, 6.81)	
Multiple pains	11.34 (8.85, 14.52)		8.57 (6.95, 10.57)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D24 Item h (walking half a mile): stratified by age group. OR (95% CI)

Independent variable	70 to 79 years		80 years and over	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
Female	1.20 (1.03, 1.39)		2.00 (1.56, 2.56)	
Lives alone	1.09 (0.93, 1.28)		0.90 (0.71, 1.15)	
School only	1.17 (0.88, 1.56)		0.99 (0.64, 1.53)	
Manual	1.31 (1.11, 1.54)		1.26 (0.96, 1.64)	
Inadequate	1.79 (1.52, 2.10)	1.41 (1.19, 1.67)	1.22 (0.97, 1.54)	
Pain elsewhere only	2.25 (1.72, 2.95)		1.40 (0.92, 2.11)	
Lower limb pain only	4.65 (3.59, 6.02)		3.27 (2.23, 4.80)	
Lower limb and low back pain	3.94 (3.00, 5.17)		2.70 (1.81, 4.04)	
Lower limb pain and pain elsewhere	5.02 (4.01, 6.30)		3.42 (2.44, 4.79)	
Multiple pains	6.64 (5.40, 8.16)		5.02 (3.61, 7.00)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D25 Item h (walking half a mile): stratified by educational attainment. OR (95% CI)

Independent variable	Further education*		School-age education	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
Female	1.16 (0.89, 1.51)		1.14 (1.05, 1.24)	
55 to 59 years	1.48 (1.01, 2.17)		1.18 (1.01, 1.38)	
60 to 64 years	2.00 (1.33, 3.00)		1.65 (1.42, 1.93)	
65 to 69 years	2.39 (1.52, 3.77)		2.13 (1.83, 2.49)	
70 to 74 years	4.49 (2.80, 7.19)		3.36 (2.86, 3.93)	
75 to 79 years	8.25 (4.91, 13.86)		5.06 (4.28, 5.99)	
80 to 84 years	14.35 (7.37, 27.91)		8.30 (6.85, 10.07)	
85 to 89 years	36.27 (16.84, 78.13)		17.50 (13.14, 23.30)	
90 years and over	57.40 (11.57, 284.70)		52.22 (30.51, 89.39)	
Lives alone	1.03 (0.75, 1.39)		1.13 (1.02, 1.24)	
Manual	1.51 (1.16, 1.97)		1.33 (1.22, 1.46)	
Inadequate	2.65 (2.03, 3.46)		1.95 (1.80, 2.12)	
Pain elsewhere only	1.41 (0.82, 2.42)		1.89 (1.60, 2.23)	1.75 (1.41, 2.18)
Lower limb pain only	5.31 (3.25, 8.70)		3.95 (3.36, 4.65)	3.65 (2.99, 4.45)
Lower limb and low back pain	6.02 (3.77, 9.61)		3.36 (2.85, 3.97)	3.39 (2.76, 4.15)
Lower limb pain and pain elsewhere	4.79 (3.12, 7.36)		5.54 (4.82, 6.38)	4.38 (3.69, 5.20)
Multiple pains	9.34 (6.44, 13.55)		8.25 (7.27, 9.36)	7.61 (6.52, 8.89)

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D26 Item h (walking half a mile): stratified by occupational class. OR (95% CI)

Independent variable	Non-manual		Manual	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
Female	1.28 (1.11, 1.49)		1.10 (1.00, 1.20)	
55 to 59 years	1.31 (1.00, 1.71)		1.19 (1.00, 1.41)	
60 to 64 years	1.86 (1.41, 2.45)		1.65 (1.39, 1.96)	
65 to 69 years	2.25 (1.71, 2.96)		2.18 (1.84, 2.59)	
70 to 74 years	4.05 (3.06, 5.36)		3.30 (2.77, 3.94)	
75 to 79 years	6.68 (4.95, 9.01)		4.88 (4.05, 5.89)	
80 to 84 years	11.46 (8.10, 16.23)		7.98 (6.43, 9.91)	
85 to 89 years	22.80 (14.20, 36.61)		17.84 (12.87, 24.72)	
90 years and over	60.75 (23.42, 157.59)		49.30 (27.11, 89.68)	
Lives alone	0.95 (0.79, 1.15)	1.18 (0.95, 1.46)	1.15 (1.04, 1.29)	
School only	1.32 (1.10, 1.57)		1.15 (0.95, 1.40)	
Inadequate	2.22 (1.91, 2.58)		2.00 (1.81, 2.21)	1.80 (1.61, 2.02)
Pain elsewhere only	2.33 (1.72, 3.14)	1.54 (1.02, 2.32)	1.69 (1.41, 2.03)	
Lower limb pain only	5.63 (4.22, 7.51)	3.37 (2.35, 4.84)	3.64 (3.05, 4.34)	
Lower limb and low back pain	5.15 (3.87, 6.85)	4.12 (2.90, 5.85)	3.15 (2.63, 3.77)	
Lower limb pain and pain elsewhere	6.68 (5.19, 8.60)	4.12 (3.01, 5.64)	4.81 (4.14, 5.59)	
Multiple pains	9.84 (7.84, 12.35)	7.12 (5.40, 9.38)	7.88 (6.88, 9.03)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Table D27 Item h (walking half a mile): stratified by perceived adequacy of income. OR (95% CI)

Independent variable	Adequate		Inadequate	
	A little or a lot[a]	A lot[b]	A little or a lot[a]	A lot[b]
Female	1.25 (1.12, 1.40)		1.05 (0.94, 1.17)	
55 to 59 years	1.24 (0.99, 1.55)	1.04 (0.73, 1.49)	1.24 (1.02, 1.50)	
60 to 64 years	1.60 (1.27, 2.01)	1.96 (1.40, 2.74)	1.77 (1.46, 2.15)	
65 to 69 years	2.57 (2.06, 3.20)	2.78 (2.01, 3.84)	1.92 (1.57, 2.34)	
70 to 74 years	3.89 (3.11, 4.88)	5.07 (3.69, 6.96)	3.06 (2.49, 3.77)	
75 to 79 years	6.49 (5.11, 8.24)	9.10 (6.61, 12.54)	4.04 (3.24, 5.03)	
80 to 84 years	11.68 (8.84, 15.43)	13.56 (9.61, 19.13)	6.38 (4.91, 8.28)	
85 to 89 years	23.49 (15.75, 35.04)	27.35 (17.98, 41.61)	14.49 (9.43, 22.28)	
90 years and over	64.85 (29.68, 141.67)	97.12 (49.90, 189.06)	22.59 (9.61, 53.08)	
Lives alone	1.09 (0.95, 1.25)		1.12 (0.99, 1.27)	
School only	1.34 (1.12, 1.60)		1.12 (0.92, 1.37)	
Manual	1.39 (1.24, 1.56)		1.30 (1.14, 1.48)	
Pain elsewhere only	1.70 (1.37, 2.10)		1.94 (1.55, 2.43)	
Lower limb pain only	3.90 (3.19, 4.77)		4.08 (3.27, 5.09)	
Lower limb and low back pain	3.40 (2.76, 4.20)		3.83 (3.08, 4.77)	
Lower limb pain and pain elsewhere	4.85 (4.05, 5.80)		5.43 (4.53, 6.51)	
Multiple pains	7.33 (6.24, 8.61)		9.18 (7.78, 10.83)	

[a] Reference group is not limited; [b] Reference group is not limited or limited a little

Appendix E Rasch model: probability of reporting difficulty with an item

Scenario 1: Person location = item location

i.e. $\beta_n = \delta_i$

Then, $\beta_n - \delta_i = 0$

Hence, $p_{ni} = \frac{e^0}{1+e^0} = \frac{1}{1+1} = 0.5$

Scenario 2: Person location two logits greater than item location

i.e. $\beta_n = \delta_i + 2$

Then, $\beta_n - \delta_i = 2$

Hence, $p_{ni} = \frac{e^2}{1+e^2} = \frac{7.389}{1+7.389} = 0.88$

Scenario 3: Person location two logits less than item location

i.e. $\beta_n = \delta_i - 2$

Then, $\beta_n - \delta_i = -2$

Hence, $p_{ni} = \frac{e^{-2}}{1+e^{-2}} = \frac{0.135}{1+0.135} = 0.12$

Appendix F Derivation of an interval-level measure of locomotor disability using items from the PF-10

Appendix F1 Fit of the five individual locomotor disability-specific PF-10 items to the Rasch model

Model: Partial Credit

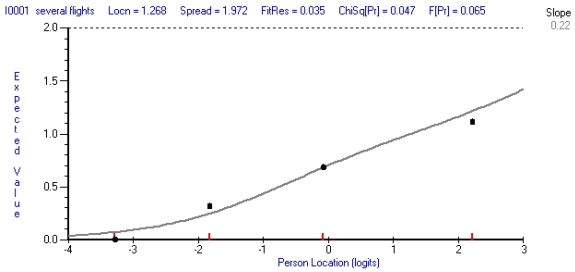
Thresholds: Ordered as expected

Table F1 Summary of Rasch model fit statistics, n=363

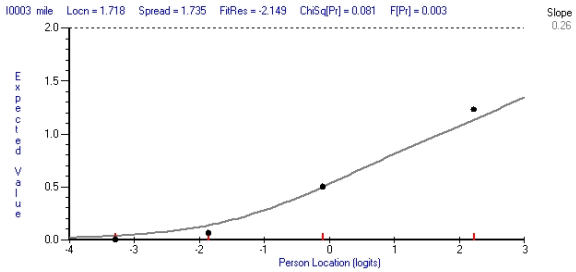
Statistic	Value in NorStOP pilot dataset
% of t-tests (95% confidence interval)	7.9% (5.0%, 10.9%)
Residual correlation (min, max)	-0.496, +0.509
Mean item fit residual (SD)	-1.072 (1.396)
Item residual	
Climb several flights of stairs	0.035
Climb one flight of stairs	0.762
Walk more than a mile	-2.149
Walk half a mile	-2.398
Walk 100 yards	-1.609
Item chi-square test (X^2 (df): p-value)	
Climb several flights of stairs	7.971 (3): 0.0466
Climb one flight of stairs	10.528 (3): 0.0146
Walk more than a mile	6.735 (3): 0.0809
Walk half a mile	12.204 (3): 0.0067
Walk 100 yards	9.989 (3): 0.0187
Item F-test (F (df_1, df_2): p-value)	
Climb several flights of stairs	2.449 (3, 208): 0.0647
Climb one flight of stairs	3.128 (3, 208): 0.0267
Walk more than a mile	4.765 (3, 206): 0.0031
Walk half a mile	8.058 (3, 205): <0.0001
Walk 100 yards	6.030 (3, 210): 0.0006
Mean person fit residual (SD)	-0.391 (0.791)
Person residuals (min, max)	–1.469, 1.702
Item-trait interaction (X^2 (df): p-value)	47.426 (15): <0.0001
Mean person location (SD)	–1.377 (3.245)
PSI	0.943

Figure F1 Extent of item misfit of items to the Rasch model

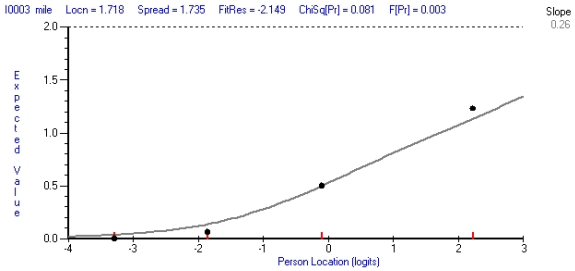
(a) Climb several flights of stairs



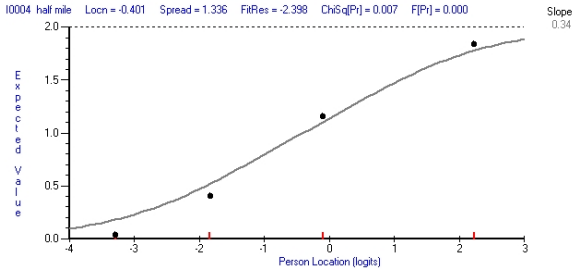
(b) Climb one flight of stairs



(c) Walk more than a mile



(d) Walk half a mile



(e) Walk 100 yards

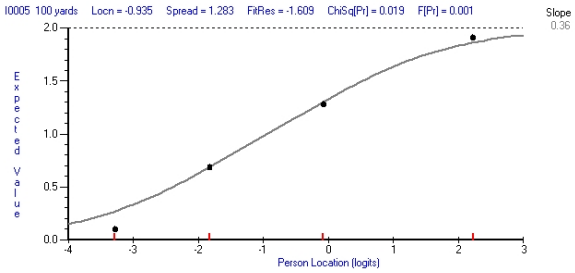
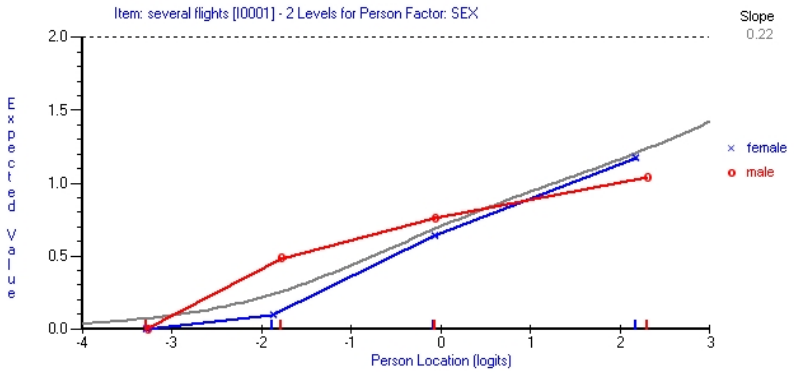


Table F2 Differential item functioning in the five PF-10 locomotion-related items by gender and age

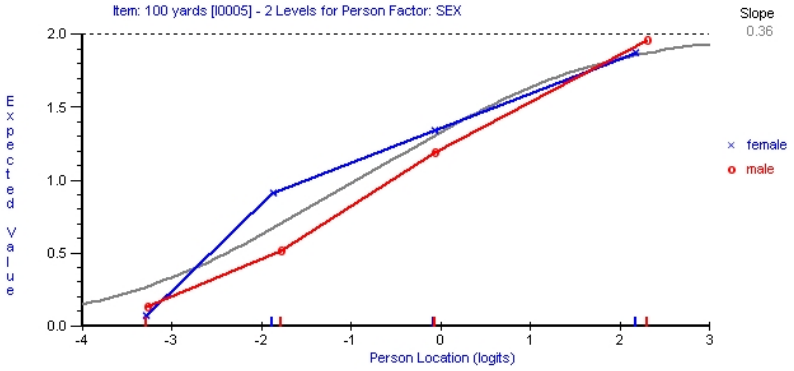
Item	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Several flights	2.046	2.803 (1)	0.0956	3.216	4.406 (3)	0.0050
One flight	1.849	2.164 (1)	0.1428	0.322	0.377 (3)	0.7699
More than mile	0.045	0.103 (1)	0.7488	0.133	0.300 (3)	0.8252
Half a mile	1.036	2.031 (1)	0.1557	-0.065	-0.128 (3)	0.9999
100 yards	1.331	2.445 (1)	0.1194	2.206	4.054 (3)	0.0079
Age						
Several flights	1.697	2.242 (2)	0.1089	0.970	1.282 (1)	0.2672
One flight	0.053	0.061 (2)	0.9405	0.705	0.816 (6)	0.5589
More than mile	0.918	2.157 (2)	0.1184	0.585	1.375 (6)	0.2265
Half a mile	0.313	0.605 (2)	0.5469	0.134	0.259 (6)	0.9552
100 yards	0.588	1.032 (2)	0.3583	0.610	1.070 (6)	0.3819

Figure F2 Extent of differential item functioning by gender

(d) Climb several flights of stairs



(i) Walk 100 yards



Appendix F2 Fit of the five individual locomotor disability-specific PF-10 items to the Rasch model: subtests for stair-climbing and walking items defined by RUMM2020

Model: Partial Credit

Thresholds: Ordered as expected

Table F3 Summary of Rasch model fit statistics, n=363

Statistic	Value in NorStOP pilot dataset
% of t-tests (95% confidence interval)	0.0% (0.3%, 0.3%)
Residual correlation (min, max)[a]	-0.837
Mean item fit residual (SD)	-0.032 (1.392)
Item residual Stair-climbing Walking	 0.952 -1.016
Item chi-square test (X^2 (df): p-value) Stair climbing Walking	 0.964 (3): 0.8098 19.879 (3): 0.0002
Item F-test (F (df_1, df_2): p-value) Stair climbing Walking	 0.591 (3, 205): 0.6216 20.769 (3, 204): <0.0001
Mean person fit residual (SD)	-0.527 (1.087)
Person residuals (min, max)	-4.075, +0.985
Item-trait interaction (X^2 (df): p-value)	20.844 (6): p=0.0020
Mean person location (SD)	–0.702 (1.857)
PSI	0.900

[a] Only one value of residual correlation as only 2 items (subtests)

Figure F3 Extent of misfit of the walking super-item to the Rasch model

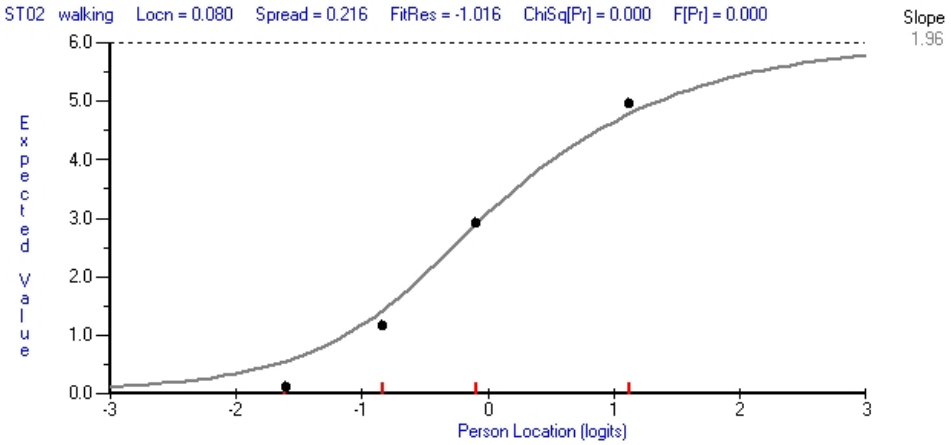


Table F4 Differential item functioning in the two subtests formed by RUMM2020 by gender and age

Sub-test	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Stair-climbing	1.533	2.683 (1)	0.1030	1.198	2.095 (3)	0.1020
Walking	0.731	2.614 (1)	0.1075	0.442	1.579 (3)	0.1956
Age						
Stair-climbing	1.513	2.546 (2)	0.0810	-0.009	-0.015 (6)	0.9999
Walking	0.372	1.298 (2)	0.2755	0.177	0.616 (6)	0.7177

Appendix F3 Fit of the two super-items to the Rasch model before combination of response categories

Model: Partial Credit

Thresholds: Disordered

Table F5 Summary of Rasch model fit statistics, n=363

Statistic	Value in NorStOP pilot dataset
% of t-tests (95% confidence interval)	0.0% (0.39%, 0.39%)
Residual correlation[a]	-0.965
Mean item fit residual (SD)	0.049 (0.846)
Item residual	
Stair-climbing	0.647
Walking	-0.549
Item chi-square test (X^2 (df): p-value)	
Stair-climbing	0.992 (3): 0.8033
Walking	18.639(3): 0.0003
Item F-test (F (df_1, df_2): p-value)	
Stair-climbing	0.499 (3, 144): 0.6834
Walking	24.291 (3, 150): <0.0001
Mean person fit residual (SD)	-0.527 (1.087)
Person residuals (min, max)	-1.561, +0.641
Item-trait interaction (X^2 (df): p-value)	19.631 (6): p=0.0032
Mean person location (SD)	–0.915 (1.888)
PSI	0.872

[a] Only one value of residual correlation as only 2 items (subtests)

Figure F4 Extent of misfit of the walking super-item to the Rasch model

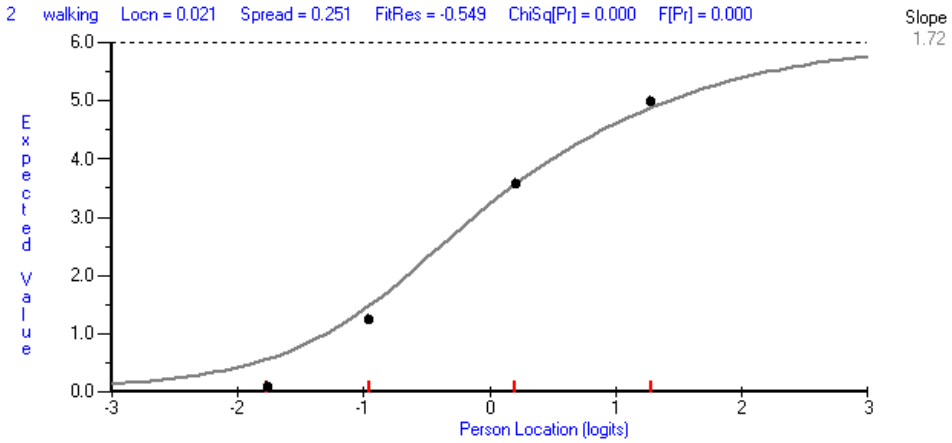


Table F6 Differential item functioning in the two super-items by gender and age

Super-item	Uniform			Non-uniform		
	Mean square	F (df)	p-value	Mean square	F (df)	p-value
Gender						
Stair-climbing	0.947	2.030 (1)	0.1564	0.925	1.983 (3)	0.1193
Walking	0.190	0.999 (1)	0.3190	0.278	1.460 (3)	0.2281
Age						
Stair-climbing	0.0818	0.167 (1)	0.6833	0.141	0.289 (3)	0.8335
Walking	0.7492	3.837 (1)	0.0520	-0.139	-0.713 (3)	0.9999

Appendix G Translations of the five PF-10 locomotor disability-related items into Welsh and Dutch

Table G1 Translation of the PF-10 locomotor disability-specific items and responses from English into Dutch and Welsh

English	Welsh	Dutch
Questions		
Climbing a several flights of stairs	Dringo sawl rhes a risau	Een paar trappen oplopen
Climbing one flight of stairs	Dringo un rhes a risau	Eén trap oplopen
Walking more than a mile	Cerdded mwy na milltir	Meer dan een kilometer lopen[a]
Walking half a mile	Cerdded sawl canllath	Een halve kilometer lopen[b]
Walking 100 yards	Cerdded canllath	Honderd meter lopen[c]
Response options		
No, not limited at all	Ydy, yn cyfyngu llawer	Nee, helemaal niet beperkt
Yes, limited a little	Ydy, yn cyfyngu ychydlg arnaf	Ja, een beetje beperkt
Yes, limited a lot	Nac ydy, ddlm yn cyfyngu arnaf o gwbl	Ja, ernstig beperkt

[a]Translated into metric system: more than a kilometre; [b]Translated into metric system: half a kilometre; [c]Translated into metric system: 100 metres