Modern psychometrics and the design and application of patient-reported outcome measures

James Twiss

PhD by publication (route 2)

A thesis submitted in partial fulfilment of the requirements of

Manchester Metropolitan University for the degree of

Doctor of Philosophy by Published Work (Route 2)

Faculty of Health, Psychology and Social Care

Manchester Metropolitan University

2015

Contents

Abstract

Creating accurate, high quality measurement with patient-reported outcomes (PRO) is a key challenge for developers. It is often the case that PRO measures fail to clearly define the constructs that they are intended to measure. Consequently, they fail to provide measurement that is valid and meaningful.

Classical test theory has been applied in the development of most outcome measures currently in use. Such psychometric approaches to PRO measure development are being superseded by more powerful item response theory (IRT) methods. The Rasch model is the one parameter form of IRT that embodies fundamental measurement requirements. Scales that produce data fitting the Rasch model provide interval level measurement, improving their power and discrimination.

The thesis argues that it is a combination of clear construct definition and application of Rasch analysis that lead to improved measurement.

The aims of the thesis are to i) describe approaches to construct definition and psychometric measurement ii) evaluate my own research in relation to these approaches iii) critically assess the contribution of the research to the field.

The thesis considers ten articles relating to the development and application of PROs. The articles in the thesis cover the following topics:

New PRO scale developments. Three articles describe developments of new measures that are based on a clearly defined construct and apply Rasch analysis in their development.

Application of PRO scales in international research. Two articles describe the adaptation of PRO scales into several additional languages. Such adaptations increase the value of the measures to international research. In addition, a minimal important difference (MID) study is described in one article. The MID estimations generated assist the interpretation of scores and sample size determination for future studies.

Evaluation of existing PRO scales. Three studies describing the evaluation of PRO measures are discussed. Weaknesses were identified in each of the scales. The Mood Disorder Questionnaire, a screening tool for bipolar disorder, was found to screen patients more effectively when the symptoms section of the scale was used without the other sections of the questionnaire. The Dermatology Life Quality Index (DLQI) and the SF-36 had several measurement limitations and were not based on clearly defined constructs.

Co-calibration of disease-specific PRO scales. A new method for combining scores from two disease-specific PROs using Rasch analysis is discussed. This method offers a means of combining PRO data from patients with different diseases that complete different disease-specific measures. This approach was possible as both measures were developed based on the same clearly defined construct and both produced data fitting the Rasch model.

The research makes a number of important contributions to the improvement of PRO measurement. The studies show that clear construct definition and application of Rasch analysis are central to improving the science. More work is necessary, particularly to understand in greater detail the needs-based model of quality of life that has been applied in the new measure development described in the thesis.

Acknowledgements

I would like to thank my supervisor, Francis Fatoye, for his positive and encouraging advice. His help was essential for keeping my writing on track. I would also like to thank my many inspiring colleagues that have contributed to the research in the thesis. Thank you to Stephen McKenna for mentoring me in the early years of my research career. Thank you also for your help in proof reading and providing challenging debate. A special thank you to all of the patients involved in the work; you have helped me understand how lives are affected by chronic illness. Finally, thank you to my parents, family and friends; without knowing it you have been a huge support to me throughout this process.

Chapter 1: Introduction

1.1 Background

For the last ten years I have been part of a research group specialising in the development of patient reported outcome (PRO) measures. In essence, these types of outcomes take the form of questionnaires and provide a means for patients to report on the impact of a disease from their own perspective. During this time I became aware not only of the importance of capturing the patients experience but also of measuring it accurately. My research interests have focussed on trying to improve the science of measurement in this area.

As scientists, we attempt to acquire and organise knowledge about phenomena into testable predictions and observations. Central to this process is the ability to create accurate and meaningful measurements. These then become the foundation of information about the phenomenon of interest and can be used to make predictions. In the physical sciences great emphasis has been made on producing accurate measurements of concepts such as temperature, mass and length (Taylor, 1991). It is vital that the same kind of rigorous approach is applied to measuring the impact of health conditions. The importance of accurate measurement is described eloquently by one of the pioneers of scientific measurement (Kelvin, 1883):

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."

Health outcomes can be measured in a variety of ways. In clinical trials, physiological measures are the most frequent form of primary outcome (Doward et al, 2010). These measures focus on assessing the physical manifestations of a disease. For example, in cardiology disease various measurements of cardiac output and blood flow can be taken to reflect the functioning of the heart. Although physiological measures can provide detailed information regarding the disease they may be invasive, expensive, time consuming and difficult to use on a regular basis. Crucially, these outcomes provide only limited perspective on a disease and as such may not accurately reflect the impact the disease or intervention has on the patient.

PRO data provide an alternative method of gaining information on the impact of a disease and offer several advantages. First, PROs allow the patient to provide their perspective on how their illness affects them. Patients' views often correlate poorly with physical assessments (Piquette et al, 2000; Jones, 2001) and differ to those of clinicians (Hewlett, 2003; Martin et al, 2009; Wehmeier et al, 2007). Patients are more likely to focus on the psychological and broader impact of an illness rather than on its physical effects (Neville et al, 2000; Doward et al, 2009a). Importantly, the broader aspects of an illness, such as participation and quality of life, can only be assessed accurately by the patient.

Various stakeholders have an interest in seeing the patient-perceived benefits of interventions including patient groups, payers, regulators, policy makers, health technology assessment bodies and clinicians (Doward et al, 2010). The ease with which PRO data can be incorporated into research studies compared with invasive physiological assessments also makes their use attractive. These factors have led to the increased use of PROs in clinical trials (Scoggins and Patrick, 2009). Despite the potential benefits, creating accurate measurements using PROs has been extremely challenging for developers and methods have lagged behind measurement in the physical sciences.

Until recently, classical test theory (CTT) psychometric approaches dominated the field of PRO development. These approaches are based on true score theory (Allen and Yen, 2002; Novick, 1996; Traub, 1997). They focus on total score level data and error associated with it. Using this approach the distances between each item in terms of the amount of construct measured is not known. The end result is an ordinal based measure with limited mathematical qualities. For example, only less powerful non-parametric statistical techniques are justified with this level of data and calculation of change scores in clinical trials is not justified (Tennant et al, 2004).

A major shift in the approach to measurement has occurred in recent years due to the application of Item Response Theory (IRT). IRT is a paradigm for the scoring, interpretation and analysis of tests (Hambleton et al, 1991). It is based on the premise that the selection of a response to an item is a probabilistic mathematical function of the amount of difficulty represented by the item and the level of trait that the person exhibits (Hambleton et al, 1991). By modelling responses to items they can be located on an underlying metric in order of severity. This contrasts with classical psychometric approaches that do not make any assumption regarding the amount of construct represented by the items. The 'item' level diagnostic information provided by IRT methods makes the approach a much more powerful method for the assessment of scale functioning than classical psychometric methods.

The Rasch model (Rasch, 1960/1980) is a simple one parameter form of IRT with particularly strong mathematical characteristics. It offers the prospect of creating scales that meet fundamental measurement requirements providing the same measurement quality observed in the physical sciences (Luquet et al, 2001; Prieto et al, 2003; Tennant et al., 2004; Waugh and Chapman, 2005; Wright, 1996; Wright and Tennant, 1996). For this reason the Rasch model is the model applied throughout the body of my research.

Despite these improvements in psychometric methods the most important component to creating accurate PRO measures is a thorough understanding and definition of the construct that is being measured. There has often been a lax approach to defining the constructs that researchers are trying to measure in PRO research. There are many examples in the field of PROs that are not based on any clear theory (McKenna, 2011; Gimeno-Santos et al, 2011). It is fundamental for accurate measurement to have a clear theoretical foundation as a starting point. This is essential to produce an outcome that is meaningful and purposeful.

1.2 Aims of the thesis

The thesis will consider ten publications and how each of these has contributed to knowledge in the field. Two important factors are considered in the work; the need for clear construct definition and the importance of rigorous psychometric approaches to measurement.

The specific aims of the thesis are:

1. Describe methods of PRO development in relation to:

a) PRO construct definition.

b) Psychometric analysis.

2. Evaluate the extent to which my own research has met these requirements.

3. Critically assess the contribution of each study to the field.

1.3 Structure of the thesis

The first two chapters will discuss the importance of clear construct definition and psychometric measurement methods in PRO development. The articles included in the thesis will be categorised into groups and presented in Chapters 4-7. Due to restrictions in copyright, only the articles published in open access journals will be included. For the remaining articles, the abstract, DOI and URL will be provided. The thesis discusses each of the articles in relation to the specified methods. It also critically reviews the research to assess how the studies could have been approached differently or improved. Finally, Chapter 8 will summarise the work, consider areas for further study and suggest how my research will develop in the future.

1.3.1 Chapter content

Chapter 1: Introduction

This chapter has provided an introduction to the topic and overview of the thesis.

Chapter 2: Construct definition

This chapter discusses the importance of clearly defining the construct that the PRO measures.

Chapter 3: Psychometric methods

In this chapter the quantitative methods for assessing the functioning of a PRO are considered. Two psychometric paradigms are discussed; CTT and IRT.

Chapter 4: Development of new PROs

In this chapter three research studies are discussed in which new PRO measures were developed.

Chapter 5: Application of PROs in international research

This chapter discusses three studies relating to the application of PROs in international research.

Chapter 6: Evaluation of existing PROs

This chapter describes three studies which evaluate the psychometric properties of existing PROs.

Chapter 7: Co-calibrating disease-specific PROs

This chapter discusses future psychometric approaches to outcome measurement in which different disease-specific measures can be combined through a process of co-calibration.

Chapter 8: Summary and Conclusions

The final chapter summarises the research, shows how the research will develop in the future and provides conclusions.

1.4 Chapter summary

Accurate measurement is fundamental to the scientific process. The quality of measurement in PRO research has lagged behind that of the physical sciences. The thesis will discuss modern psychometric methods that are necessary to provide high quality measurement. It will focus on the importance of clear construct definition and the application of Rasch analysis. The thesis presents 10 articles relating to PRO development and application. Each article will be evaluated based on their contribution to the field and the quality of the methods employed.

Chapter 2: Defining the construct

2.1 Introduction

PROs are designed to measure constructs that are not directly observable (commonly described as latent variables). The constructs can cover a broad range of health outcomes. These include symptoms, functional limitations, health-related quality of life (HRQL), well-being, participation and satisfaction with care. Whatever the type of outcome that a PRO assesses it is essential that a rationale and explanation of the construct is provided. This then forms a guide for the PROs content so that items representing the construct of interest can be selected. This chapter discusses the importance of clearly defining the construct and considers three approaches to construct definition. Each provides a different perspective in the construct definition process. Although there is a degree of overlap and commonality between the approaches it is not possible to explore this in detail in the present thesis. Instead, each will be discussed separately and their strengths and weaknesses considered.

2.2 Background

Several articles have been published detailing standards for the development of PROs (Reeve et al, 2013; Erickson et al, 2009; Magasi et al, 2011; Revicki et al, 2000; Revicki et al, 2007; Rothman et al, 2009; Snyder et al, 2011; Turner et al, 2007; US Food and Drug Administration, 2009). Most of these guidance documents identify the importance of having a clear conceptual basis for the PRO. Despite the existence of such guidelines a large number of outcome measures have been developed and continue to be developed without a clearly defined construct. For example, Nixon et al (2013) evaluated twenty six PROs available for epilepsy and concluded that none of the available measures had a clear conceptual basis.

The consequences of failing to define the PRO construct clearly are serious (Rothman et al, 2007). Without a clear definition the validity of the PRO must be questioned. Failure to define the construct properly often leads to the grouping of items that represent different kinds of outcome. For example, a PRO claiming to measure overall HRQL may confound items measuring symptoms (e.g. pain) with others measuring functioning or social impact. In the absence of clear construct definition erroneous interpretation of scores is likely as it is not clear what the scoring represents. This makes it difficult to assess the effectiveness of different interventions. Without proper construct definition accurate and purposeful measurement is not possible. Some of the approaches to PRO construct definition are examined in the next section.

2.3 Approaches to construct definition in PRO measurement

Three approaches to defining the construct that a PRO measures will be discussed in this chapter. These are:

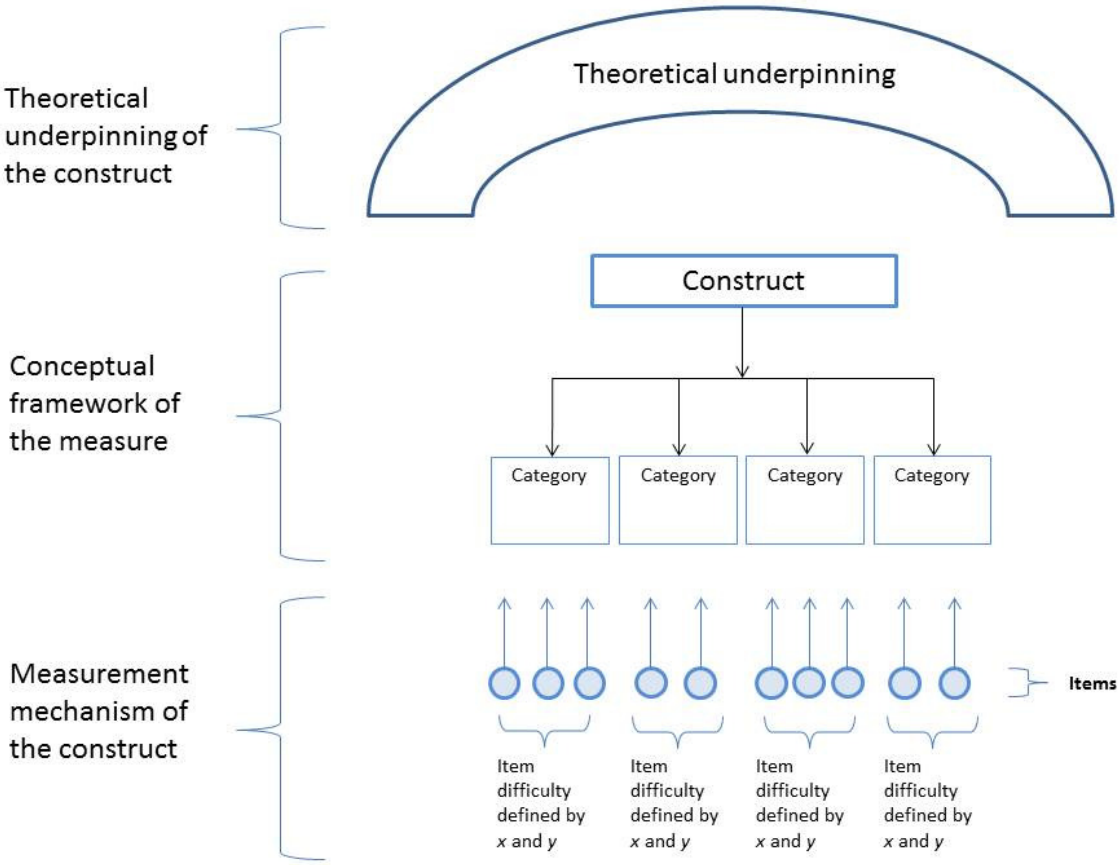
- Theoretical underpinning: A theoretical grounding for the construct of interest should be provided. This places the measure within the context of a larger body of explanatory work.

- Conceptual framework of the PRO: This approach explains the structure of the PRO and shows the relation between items, domains and the overall construct. It is usually organised in the form of a figure.

- Measurement mechanism of the PRO: This is an approach to construct definition whereby the underlying mechanism of the measure is understood so that items can be manipulated to represent varying levels of the construct of interest.

Each approach covers a slightly different component of construct definition. Figure 2.1 shows how each of the approaches may relate in the construct definition process.

Figure 2.1: Approaches to defining the PRO construct



2.4 Theoretical underpinning

The constructs to be measured should be embedded in a larger body of theoretical work. This gives each construct a greater degree of explanation and allows the construct to be understood relative to other variables important in the patients' disease.

Scientific theories have been described as nets cast to catch what we call 'the world': to rationalise, to explain, and to master it (Popper, 1963). They are developed through scientific methods including hypothesis generation and testing, deductive and inductive logic and parsimony (Gauch, 2002). They are used to represent scientific knowledge.

One of the most influential explanations of the properties of scientific theories has been provided by Popper (1959). He identified the following characteristics:

1. Theories should include causal explanations; from these explanations it is possible to make specific predictions.

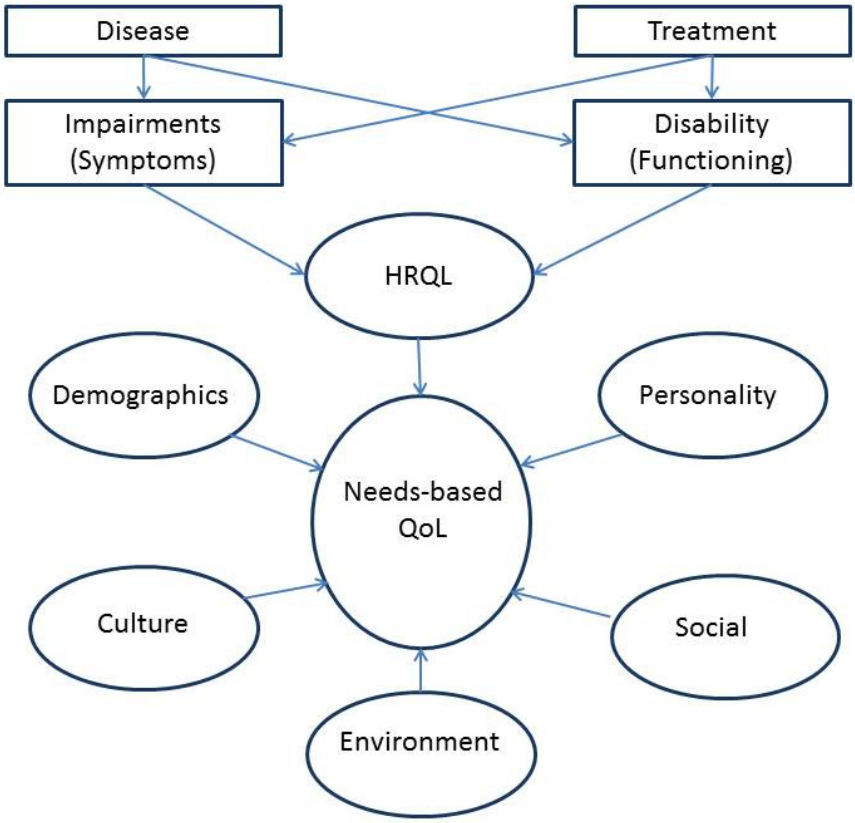
2. A theory should have the property of falsifiability; the specific predictions that the theory includes should be testable and therefore falsifiable.

The most important aspects of a scientific theory are its testability and falsifiability (Popper, 1963). This allows the content to be scrutinised scientifically. A theory that is not refutable cannot be considered a true scientific theory. More specific theories lead to clearer predictions and more testable content.

Placing a construct in a wider explanatory theory considerably strengthens the conceptual foundation of a construct. The needs-based approach to quality of life (QoL) provides an example of how this can be achieved (Hunt and McKenna, 1992). This is the approach used in some of my own research presented in the following chapters. It is built on theoretical work on human needs (Maslow, 1970; Max-Neef et al, 1991; Kenrick et al, 2010). According to the theory QoL is high when needs are fulfilled and low when they are not. Relevant needs can be seen to fall into several different categories including those related to safety and security, socialisation, affection, esteem, cognition and personal development. Different illnesses impact on patients' ability to meet their needs in different ways.

The needs-based approach also defines and explains the relation with other constructs relating to the impact of disease. QoL is clearly distinguished from the constructs of impairments and activity limitations. These latter types of outcome may influence QoL but are only important to the patient insofar as they prevent need fulfilment. Figure 2.2 shows a simplified model of the interrelations between impairments, functional limitations and needs-based QoL (McKenna and Doward, 2004). It also shows how other factors such as personality and culture may influence overall QoL.

Figure 2.2: Influences on QoL

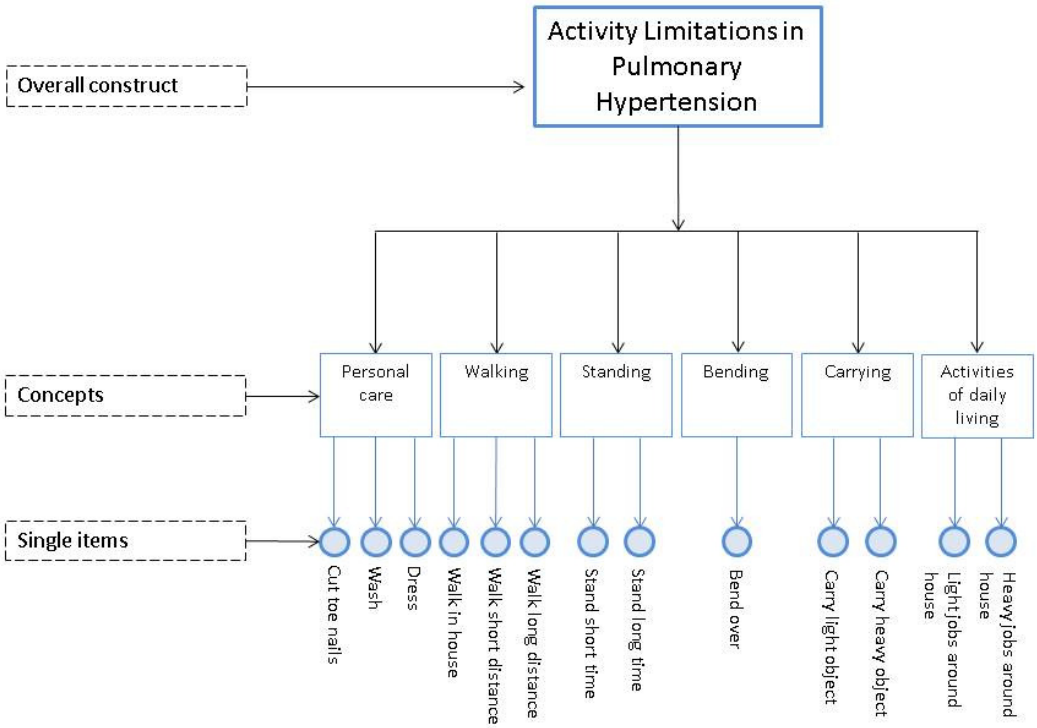


2.5 Conceptual frameworks

In addition to explaining the wider theoretical context for a construct it is also important to identify what a PRO actually measures in terms of content. A common approach to construct definition is to use conceptual frameworks for this purpose. The conceptual framework should provide an explanation of the concepts measured and the relations between the concepts. It should also show how each item relates to the concepts. The conceptual framework is usually represented figuratively (FDA, 2009; Erickson et al, 2009).

Erickson et al (2009) have attempted to guide the development of conceptual frameworks likening them to hierarchies of increasing complexity. At a very basic level there are single items. These can then be grouped into specific or generic families at a higher order. Above this level are more complex aggregate or compound concepts composed of multiple families. This PRO concept taxonomy provides the structure for a PRO. Figure 2.3 below shows an example conceptual framework based on the Cambridge pulmonary hypertension outcome review (CAMPHOR) activities limitations scale (McKenna et al, 2006). Activity limitation is defined by six different kinds of activities (e.g. walking). Below this level single items are used to measure the activity kinds (e.g. walking a short distance; walking up a slight incline). This approach clearly represents how the different kinds of activities relate to the overall construct.

Figure 2.3: Conceptual framework for activity limitations in pulmonary hypertension



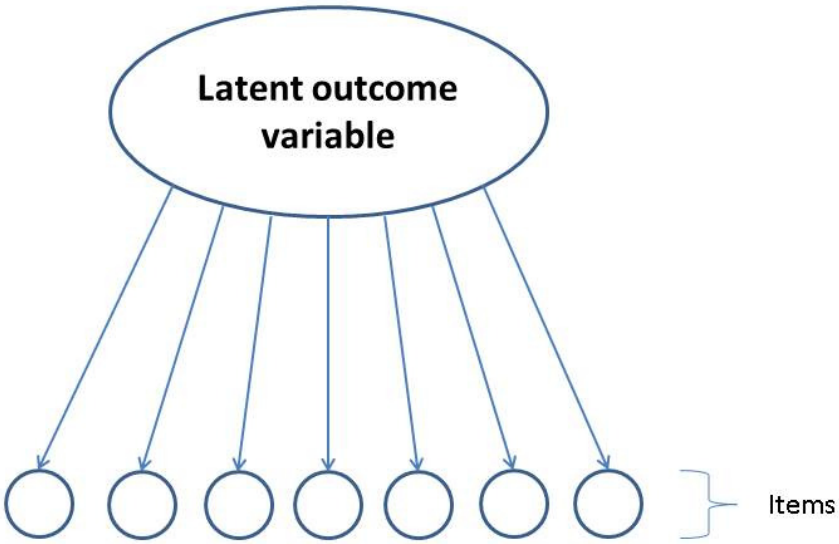
Conceptual frameworks are primarily concerned with explaining the inter-relations between items and domains of a PRO. Although they are useful in guiding PRO development they are also limited in their scope as they are descriptive rather than explanatory and may leave some parts of the construct unsolved (Donatti et al, 2008). For example, a HRQL framework may identify several different domains and how items relate to the domains. However, it may fail to explain why the domains have been chosen and how these relate more broadly to the persons health. Without this further explanation ill-conceived domains may be chosen that lack validity.

2.6 Measurement mechanism of the construct

Recent research has suggested that it is important to explain exactly 'how' a measure works and why the items represent different levels of the construct of interest (Stenner et al, 2013). To do this it is necessary to understand the measurement mechanism of the construct.

The approach most frequently used in health outcome measurement assumes that the items in a scale are manifestations of the construct (latent variable) (Stenner et al, 2008). When changes occur in the latent variable it is assumed that this will be shown in the item responses. This principle is shown in Figure 2.4. The problem with this approach is that the items may appear to be related to the construct but it is not clear exactly how they are or why the items represent different levels of the construct. To achieve this a causal explanation of the mechanism is needed.

Figure 2.4: General psychometric explanation for a latent variable



Items: descriptions of observable manifestations

Stenner et al (2013) argue that the construct should be explained by a specification equation. For example, in the physical sciences force is defined by acceleration and mass. By manipulating one of the variables a predictable change in force should be observed. In relation to outcome measurement in the social sciences, it is argued that manipulations of specific aspects of items should result in predictable changes in the level of construct measured. By finding these aspects it is possible to provide a measurement specification equation similar to that in the physical sciences. In the absence of this kind of specification equation the construct cannot be fully understood. Measures without an explanatory measurement equation have been likened to 'black boxes' where a construct may be understood and explained loosely but not fully.

These ideas have been shown clearly in a body of work leading to a model of reading ability (Stenner et al, 2013). The measurement specification equation has been defined by the authors by the two variables text complexity and sentence length. An increase in one or both of these variables will increase the difficulty of the text. Using this method all texts can be graded with a 'lexile' score. This work has led to widespread adoption of this scoring system for assessing children's progression with reading ability in the US.

The application of this approach has so far been limited to a few areas of research by the original authors. The approach has the potential to markedly improve the precision of measurement in health research. Despite the potential no outcomes are available that have adopted this approach in health research. However, it has been included in the thesis to provide an alternative perspective and to help evaluate the research included.

2.7 Chapter summary

This chapter has discussed three approaches to construct definition. Each of the approaches focuses on a different aspect of construct definition. Scientific theories provide in depth explanations about the nature of constructs. Embedding the construct definition of a PRO within a scientific theory strengthens it considerably. Conceptual frameworks provide useful and simple explanations of the inter-relations between items, domains and the constructs of a PRO. An alternative approach to this is to identify a specification equation for the construct so that it is possible to explain exactly how the measure works.

Chapter 3: Psychometric methods

3.1 Introduction

In this chapter quantitative methods for assessing the functioning of a PRO will be discussed.

Two main psychometric paradigms are available for this purpose:

1. Classical test theory analysis

2. Item response theory

3.2 Classical test theory

3.2.1 Introduction

Until the 1990's classical test theory (CTT) was the dominant paradigm in the statistical evaluation of PROs. This approach is based on true score theory (Allen and Yen, 2002; Novick, 1996; Traub, 1997). The approach assumes that a person has a 'true score' or an accurate score that represents their real level/ability on a test. However, this true score is also obscured by error that is inherent in the test. The observed score is expressed in the following way (Kline, 2005):

Observed score = True score + error

Classical psychometric approaches attempt to explain the relation between these variables. In terms of scale construction, important methods within this framework include assessments of reliability and factor analysis (Allen and yen, 2002; Pett et al, 2003). Both methods are based on correlational analyses. Reliability is defined as the degree of consistency of a scale. Various forms of reliability assessment are available and their use within scale development aims to reduce the level of unexplained error inherent in a scale. Factor analysis is used to explore the inter-relations between items in order to group the items into smaller sets of explanatory dimensions or factors.

A brief description of some of the classical psychometric approaches to scale development is provided below.

3.2.2 Reliability

Approaches to the assessment of reliability fall into two main categories: internal reliability and test-retest reliability (reproducibility).

3.2.2.1 Internal reliability

Internal reliability is usually assessed using Cronbach's alpha coefficients (Cronbach, 1951) as a means of assessing the extent to which the items in a scale are inter-related. Traditionally, a low alpha (below 0.7) is taken to indicate that the items do not work together to form a scale (Nunnally and Bernstein, 1994).

However, this form of assessment has come under criticism for several reasons. It is well known that it can be artificially inflated by selecting items that are homogeneous (Hattie, 1985; Barchard and Hakstian, 1997; Raykov, 1997). This means that items at the extreme of a scale, which may be important for increasing the measurement range and precision of the instrument, may be discarded due to lower item-total correlations. Cronbach's alpha can also be increased by simply increasing the number of items in a scale (Nunnally and Bernstein, 1994; Streiner, 2003). These findings indicate that this form of reliability should be interpreted with caution as it is prone to bias.

3.2.2.2 Test-retest reliability

The test-retest reliability of a measure is an estimate of its reproducibility over time when no change in condition has taken place. It is assessed by correlating scores on the scale obtained on two different occasions. A high correlation indicates that the instrument produces a low level of measurement error.

Various statistics are used for the correlation analysis including Pearson correlation (Pearson, 1895), Spearman's rank correlation (Spearman, 1904) and intra-class correlation (ICC) (Koch, 1982). Some have argued that ICC is more appropriate for the assessment of reliability as, unlike Pearson correlation, it takes into account both within-subject change and systematic change in mean (Lexell and Downham, 2005). However, Spearman's rank correlation is the most appropriate for PRO data as they are ordinal in nature.

There is no widely accepted consensus on what the minimum level of reliability of a test should be. It has been argued that a minimum Pearson value of 0.85 should be used (Streiner and Norman, 1989). This is because the level of explained variance in scores is the square of the correlation value. This means a correlation value of 0.85 represents 72% explained variance. Conversely 28% of variance in scores is unexplained. The level of unexplained variance increases sharply as the correlation coefficient decreases.

3.2.3 Factor analysis

Factor analysis includes a group of statistical methods that are used to identify the relations between a set of variables or questionnaire items in order to group them into a smaller number of explanatory domains (Nunnally and Berstein, 1994). It is often used in instrument development in order to investigate the internal structure of the construct of interest.

There are two basic forms of factor analysis: exploratory and confirmatory factor analysis (Pett et al, 2003). Exploratory factor analysis is used when there is uncertainty regarding a construct. It attempts to find relations between items in order to define the construct. Confirmatory factor analysis is used when there is an existing hypothesis regarding the structure of a construct. It is used to assess how well a set of data fit this hypothesized construct.

The initial steps in a factor analysis are performed using Pearson product moment correlations. Many of the assumptions of the Pearson correlation are therefore applicable to factor analysis including the requirement for large sample sizes, normality of distributions and linear relationships between items (Pett et al, 2003).

Factor analysis has been criticised on several grounds. It involves a series of different statistical approaches and there is no consensus on which method is the most appropriate to use. For example, in the SPSS statistical package (IBM Corp, 2011), for exploratory factor analysis there are seven options for the 'extraction method', six options for the 'rotation', and also 'covariance' and 'correlation matrix' approaches leading to 84 different options for the analysis (Christenesen et al, 2012). Different approaches may lead to different results. Therefore, there is a large degree of subjectivity and 'artistry' involved in the method. Furthermore, misinterpretation of results is common (Pett et al, 2003).

The major limitation of the approach though is that factor analysis provides little information regarding the functioning of the measurement properties of a scale (Christensen et al, 2012; Wright, 1996). It does not inform on the hierarchy of items in terms of their difficulty. Neither does it inform on whether a scale meets the requirements of interval level measurement. No formal assessments of these properties are available within the factor analysis framework. In fact one of the assumptions of the approach is that there is already a linear relation among the items.

3.2.4 Limitations of CTT

The main advantages of CTT are that it is based on relatively weak assumptions and it requires little mathematical knowledge on the part of the user (De Champlain 2010). This means the methods are easier to use and more accessible to developers of new PROs.

Several limitations of the methods employed in CTT have been discussed above. In addition to these, CTT has been criticised on several levels due to its weak underlying assumptions (Petrillo et al, 2015; Xitao, 1998; Hambleton et al, 1991). True scores in the population are assumed to be measured at the interval level and normally distributed. However, PROs provide ordinal measurement and data are often not normally distributed. CTT also produces findings that are both sample and scale dependent. This is problematic as the measurement performance of a scale can be distorted by the sample it is drawn from.

CTT methods of PRO development provide ordinal level measures that should not be used in parametric assessments (Tennant et al, 2004). Parametric analyses provide more powerful and complex analyses than their non-parametric counterparts. They include the use of change scores in clinical trials, regression and analysis of variance statistics. Unfortunately the data requirements are usually ignored by most researchers using PRO measures leading to unknown consequences for study results.

3.3 Item response theory

3.3.1 Introduction to item response theory

Item response theory (IRT) includes a group of models that are concerned with the design, analysis and scoring of tests. Mathematically they are more sophisticated than their CTT counterparts and are primarily focused on item rather than test-level information. Unlike simpler CTT approaches, IRT does not assume that each item is equally difficult. Instead it views each item as representing a different level of the construct. IRT models the response of patients of a given ability to an item of a given difficulty (or severity of health outcomes).

In IRT the probability of a given response is a mathematical function of the person and item parameters. The person parameter is the latent trait being measured (for example functional disability). Several different parameters may be included at the item level. The number of parameters here defines the type of IRT approach (Chong, 2013). Item parameters may include item difficulty, discrimination and guessing. Correspondingly these would represent one, two or three parameter models.

The method adopted in my own research is the Rasch model (Rasch, 1960/1980). The Rasch model is a simple logistic one parameter IRT model. It differs importantly from other IRT models in terms of its approach to measurement (Andrich, 2004). Other IRT models focus on attempts to find a model that fits the data. In contrast the Rasch approach focuses on whether data fit the Rasch model. The Rasch model has a number of special properties that make it particularly strong in terms of measurement. This is discussed in the following section.

3.3.2 The Rasch model

The Rasch model may be conceived of as a probabilistic version of the Guttman scale (Guttman, 1950). According to the Rasch model, the probability that an individual will respond in a certain way to a particular item is a logistic function of the relative distance between the item location parameter and the person location parameter. This difference governs the probability of the expected response for a person, of a given ability, to an item of a given severity. In other words, the probability of a person affirming an item depends on how much a person is affected and the severity of the item.

The simplest Rasch model is the dichotomous model and can be formalised in the following way (Rasch, 1960/1980):

$$p_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Where $p_i(\theta)$ is the probability that persons with ability (severity) θ will affirm item *i*, and *b* is the item difficulty parameter.

3.2.3 Advantages of the Rasch model

The Rasch model was selected in my own research due to its special mathematical properties, in particular, its satisfaction of fundamental measurement requirements (Tennant et al, 2004; Wright, 1997). When items fit the model they possess a number of characteristics including criterion-related construct validity, unidimensionality, additivity, specific objectivity and sufficiency. The characteristics of specific objectivity and sufficiency are requirements of fundamental measurement and the Rasch model is the only form of IRT that achieves this.

For a measure to achieve specific objectivity in relation to the latent construct, comparison between two persons should be independent of the test or particular set of items selected from the test that are chosen for the comparison (Stenner, 1994). The property of sufficiency means that the person score on the scale contains all the available information within the specified context about the individual (Linacre, 1992).

The special properties of the Rasch model means it can be considered a gold standard for measurement in the health sciences. When data fit the model interval measurement can be achieved.

3.3.2.4 Fit to the Rasch model

The study of Rasch model fit is a process where no individual fit statistic is either necessary or sufficient to confirm the model (Andrich, 1988; Andrich and Sheridan, 2009). Therefore, the final interpretation of fit is a collective one based on several indices. If satisfactory fit is achieved across this range of different indices then a scale can be considered to show fit to the Rasch model. Different statistical packages provide different fit statistics. The fit statistics described below are available using RUMM2030 package (Andrich and Sheridan, 2009).

3.3.2.4.1 Person separation index (PSI)

Internal reliability is analysed using the PSI. The PSI is indicative of the power of the items to distinguish between respondents. A PSI score of 0.70 is the minimum acceptable level.

3.3.2.4.2 Item level fit and overall fit

Individual item fit statistics are investigated via Chi^2 fit statistics. A significant Chi^2 fit statistic ($p < 0.05$ (Bonferroni adjusted)) indicates a significant deviation from model expectations. Individual item fit residuals are also investigated. These fit residuals are the standardised sum of the squared residuals. These should fall within ±2.5 if all individuals respond in the anticipated way. High negative residuals are indicative of item redundancy and high positive residuals multidimensionality.

The overall scale fit to the model is examined by reference to several indices. An overall item-trait interaction Chi^2 fit value is calculated based on item level statistics. A significant Chi^2 statistic ($p < 0.05$) indicates that there is a real deviation of the scale from the expected pattern and a lack of fit to the Rasch model. Overall fit of the data is also investigated via Item and Person interaction statistics. These assessments measure the extent to which observed item and person estimates deviate from the expected. The mean location of the items is always anchored at 0. Within this function both Person fit residual and Item fit residual statistics are transformed by RUMM to approximate Z-scores; representing a standardised normal distribution. When the data fit the model the overall distribution statistics for Item fit and Person fit should have a mean of approximately 0 and a standard deviation of approximately 1.

3.3.2.4.3 Differential item functioning (DIF)

A requirement of the Rasch model is that items should be invariant across groups. This is investigated through tests of DIF. DIF represents instability in the order of severity of items and indicates that the scale may not work in the same way in different sub-groups of individuals (such as those defined by age or gender) that share the same level of trait being measured (Holland and Wainer, 1993). An ANOVA of standardised residuals is carried out to examine DIF by relevant groups.

3.3.2.4.4 Targeting of the measure

Targeting of items to the respondents can also be assessed by examining person-item distribution graphs. These show the ordering of both persons and items on the same logit scale and indicate whether the items in the scale are well matched to the respondents. This can help to identify where additional items are required to improve measurement along the latent-trait.

3.3.2.4.5 Functioning of the response options

The Rasch model allows formal assessment of the functioning of the response options of the measure. The probability of each response being endorsed should increase logically as the level of trait exhibited by the persons increases. The point between two adjacent categories (where the probability of endorsing either reaches 0.5) is called the threshold. For data to fit the model the threshold points should be correctly ordered so that persons are more likely to select the responses in the logical order. Participants with higher levels of trait would also be more likely to select the response categories representing high level of trait (and vice versa). Disordered response thresholds occur when participants have problems discriminating between different response categories.

3.3.2.4.6 Local dependency

One of the requirements of the Rasch model is the local independence of items. Local dependency occurs when items are too closely related such that the response to one item has too strong an influence over answers to another item (Tennant and Conaghan, 2007; Baghaei, 2008). The practical impact of this is a spreading of the Rasch estimates as they become slightly more predictable (Baghaei, 2008). Correlation of the residuals (after the underlying trait is conditioned out) should therefore be close to 0.

3.4 Chapter summary

CTT has been the dominant force in the statistical evaluation of PROs. Although simple and practical to use these methods have weak underlying assumptions. In addition, the methods provide only limited, ordinal level data.

IRT is built on more robust assumptions than CTT. The Rasch model is a one parameter form of IRT with particularly strong mathematical properties. When data fit the Rasch model then fundamental measurement is achievable. For this reason the Rasch model has been given preference in the work presented in this thesis.

Chapter 4: Development of new PROs

4.1 Introduction and articles

This chapter discusses three articles describing the development of new PROs. Each study describes the development of a new disease-specific PRO and used the same development methods. The conceptual basis and measurement approach used in each study is discussed. Finally the methods used are evaluated in terms of their suitability.

4.1.1 Article 1: The development of patient-reported outcome indices for multiple sclerosis (PRIMUS)

Doward LC, McKenna SP, Meads DM, Twiss J, Eckert BJ. The development of patient-reported outcome indices for multiple sclerosis (PRIMUS). Mult Scler. 2009 Sep;15(9):1092-102. doi: 10.1177/1352458509106513.

BACKGROUND: Complex diseases such as multiple sclerosis (MS) present dilemmas over the choice of patient-reported outcome measures as no single scale can inform on all types of MS impact from the patient's perspective. OBJECTIVE: To develop an outcome tool, the Patient-Reported Indices for Multiple Sclerosis (PRIMUS), to assess MS symptoms, activities, and quality of life.

METHODS: PRIMUS content was derived from qualitative interviews with UK MS patients and checked by clinical experts. Semi-structured cognitive debriefing interviews assessed scale face and content validity. PRIMUS scaling properties, reliability, and construct validity were assessed by a test-retest postal survey.

RESULTS: Cognitive debriefing interviews (n = 15) demonstrated scale clarity, relevance, and comprehensiveness. The postal survey was completed by 135 patients with MS. After removal of misfitting items and those exhibiting differential item functioning, all scales fitted the Rasch model, confirming unidimensionality. For all scales, test-retest reliability exceeded 0.80. Scale scores were related to perceived MS severity, general health, and symptoms of depression. Moderate correlations were observed between PRIMUS and Nottingham Health Profile scores.

CONCLUSIONS: Clinicians and researchers can have confidence in scores obtained by respondents on the PRIMUS. The PRIMUS will aid the assessment of the impact of MS from the patient's perspective.

The article is available via: http://msj.sagepub.com/content/15/9/1092.long

4.1.2 Article 2: The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS)

Meads DM, Doward LC, McKenna SP, Fisk J, Twiss J, Eckert B. The developmentand validation of the Unidimensional Fatigue Impact Scale (U-FIS). Mult Scler.2009 Oct;15(10):1228-38. doi: 10.1177/1352458509106714.

BACKGROUND: The multidimensional assessment of fatigue is complicated by the interrelation of its multiple causes and effects.

OBJECTIVE: The purpose of the research was to develop a unidimensional assessment of fatigue (U-FIS).

METHODS: Data collected with the Fatigue Impact Scale (FIS) were subjected to Rasch analysis to identify potential problems with the scale. Additional items for the U-FIS were generated from interviews with UK MS patients. The U-FIS was tested for face and content validity in patient interviews and included in a validation survey to determine dimensionality (Rasch model), reliability and validity.

RESULTS: The original FIS was not unidimensional when subscale items were combined. The modification of the FIS and addition of a number of items allowed the development of a 22-item unidimensional scale (U-FIS) that was reliable (Cronbach Alpha = 0.96; test-retest = 0.86,) and valid given correlations with the Nottingham Health Profile and ability to distinguish between MS severity groups. There was no significant difference in U-FIS scores according to MS type.

CONCLUSION: It is valid to conceptualize the functional impact of fatigue as unidimensional. The U-FIS is a reliable and valid questionnaire that will allow the measurement of this construct in clinical studies.

The article can be accessed via: http://msj.sagepub.com/content/15/10/1228.long

4.1.3 Article 3: Development and validation of the living with chronic obstructive pulmonary disease questionnaire

McKenna SP, Meads DM, Doward LC, Twiss J, Pokrzywinski R, Revicki D, Hunter CJ, Glendenning GA. Development and validation of the living with chronic obstructive pulmonary disease questionnaire. Qual Life Res. 2011 Sep;20(7):1043-52. doi: 10.1007/s11136-011-9850-6.

PURPOSE: Available patient-reported outcome (PRO) measures for chronic obstructive pulmonary disease (COPD) focus primarily on impairment (symptoms) and activities (functioning). The purpose of the study was to develop a patient-based PRO measure for COPD that captures the overall everyday impact of living with COPD from the patient's perspective.

METHODS: LCOPD items (Living with COPD Questionnaire) were generated from qualitative interviews in the U.K. and focus groups in the U.S.A. The draft measure was tested for face and content validity in both countries. Item reduction and testing for reproducibility and construct validity was conducted via Rasch and traditional psychometric analyses.

RESULTS: The draft LCOPD was found to be relevant and acceptable to patients in the U.K. (N = 19) and U.S. (N = 16). Application of Rasch analysis to data collected in validation studies (n = 162 in the U.K. and 145 in U.S.) identified a 22-item scale that measured a single construct in both countries. Psychometric analyses indicated that this version was internally consistent and reproducible. Scores on the measure were related as expected to clinician ratings of disease severity and patient ratings of COPD severity and general health.

CONCLUSIONS: The LCOPD is a new measure examining the everyday impact of living with COPD. It demonstrates good scaling properties and may prove valuable in understanding treatment benefits.

The article cab be accessed via: http://link.springer.com/article/10.1007%2Fs11136-011-9850-6

4.2 Description of articles

The Patient Reported Outcome Indices for Multiple Sclerosis (PRIMUS; Doward et al, 2009b) was developed due to the lack of high quality measures available for the assessment of the holistic impact of MS from the patients' perspective. Other available outcome measures such as the Multiple Sclerosis Quality of Life questionnaire (MSQoL-54; Vickrey et al, 1995) and Multiple Sclerosis Impact Scale (MSIS-29; Hobart et al, 2001) predominantly measure symptoms and functioning. The PRIMUS contains three scales measuring; symptoms, activity limitations and needs-based quality of life (QoL). Each scale showed good fit to the Rasch model. In addition, all scales had adequate levels of internal reliability (Cronbach's alpha >0.70) and test-retest reliability (Spearman Rank >0.80). All scales distinguished between self-perceived severity and symptoms of depression providing evidence of construct validity.

The Unidimensional Fatigue Impact Scale (U-FIS) was developed as a unidimensional measure of the impact of fatigue from the patients' perspective (Meads et al, 2009). Fatigue is one of the main symptoms of Multiple Sclerosis and has a major impact on the patient's life (Multiple Sclerosis Society UK, 1997; Freal et al, 1984; Multiple Sclerosis Council for Clinical Practice Guidelines, 1998; Krupp et al, 1988). It has often been conceptualised as a multidimensional construct including sub-components such as motor and cognitive fatigue (Schwid et al, 2002; Trojan et al, 2007). However, attempts to define fatigue as a multidimensional construct have been challenging due to the multiple causes of fatigue and the inter-relations between the different 'types' of fatigue (Penner et al, 2007).

Measures available for assessing fatigue include the fatigue severity scale (Krupp et al, 1989), The Multi-dimensional Fatigue Inventory (Smets et al, 1995), and the Fatigue Impact Scale (FIS; Fisk et al, 1994). Each of the available measures conceptualises fatigue as a multidimensional construct and were developed without the benefit of modern IRT methods. The aim of this study was to investigate whether a unidimensional measure of fatigue specific to multiple sclerosis patients could be developed. The content of the new measure was based partly on the content of the FIS and also on patient interviews. The final scale fit the Rasch model supporting the unidimensional structure of the measure. Adequate internal reliability (Cronbach's alpha=0.96), test retest reliability (Spearman's rank=0.86) and construct validity (scores significantly related to self-perceived severity and current MS flare up) was also observed.

The Living with Chronic Obstructive Pulmonary Disease (LCOPD) Questionnaire (McKenna et al, 2011) was developed as a disease-specific needs-based QoL measure. A comprehensive review identified several generic and disease-specific outcome measures for COPD (McKenna et al, 2011). However, very few of these measures covered issues related to the emotional or quality of life impact of the disease and, where they did, they were covered by only a handful of items. Due to this, it was decided to develop an outcome measure for this purpose. The measure was developed simultaneously in the UK and US. The final measure had good fit to the Rasch model in both centres. In addition, good internal reliability (Cronbach's alpha in UK and US = 0.92) and test-retest reliability (Spearman's Rank UK = 0.89; US=0.83) were observed. The scale also distinguished significantly between self-perceived severity and clinician-perceived severity.

Samples of the PRIMUS, U-FIS and LCOPD are provided in Appendices 2-4. The full measures are held under copyright and can be obtained from gr@galen-research.com.

4.3 Methodology

In each of the three studies appropriate ethics approval was sought. Informed consent was obtained from all of the participants in the studies. All data obtained from the participants was anonymised.

The development process for all the measures involved three stages:

1. Qualitative interviews and analysis to generate item content.

2. Item reduction and assessment of the psychometric properties of the scales.

3. Cognitive debriefing interviews.

4.3.1 Qualitative interviews

In the PRIMUS and U-FIS studies thirty relevant patients were included in the qualitative interviews. In the LCOPD study thirty patients were included in one-to-one interviews in the UK and 14 patients attended two focus groups in the US. The qualitative data formed the basis of the content for the PROs. As PROs are used to inform on the patient-perceived impact of an illness it is fundamental that the patient is involved.

Content analysis was conducted by the research team as a whole. The transcripts were first read by two researchers to extract issues relating to the constructs. These were then collated and coded thematically by the research team. The coding was guided by the relevant conceptual background for each measure. Finally, the themes identified were reviewed by the research groups and harmonised.

4.3.2 Item reduction and assessment of psychometric properties

A test-retest psychometric survey was conducted in each of the studies. This involved relevant patients completing the draft measures at two time points two weeks apart. In addition, patients also completed a comparator measure and questions about their overall health and disease. The purpose of the surveys was to identify scales that were unidimensional with good measurement properties. Item reduction was conducted using Rasch analysis. In addition, internal reliability, test-retest reliability and construct validity were assessed.

4.3.3 Cognitive debriefing interviews

Cognitive debriefing interviews were conducted to test the content and face validity of the draft measures. These interviews were designed to assess the clarity, relevance and comprehensiveness of the measure. The measures were completed by relevant patients in the presence of one of the developers and the interviewees asked about the ease of completion and the appropriateness of the instructions, items, and response formats. Items found to be problematic were considered for removal.

4.4 Evaluation

The development of the three measures was successful and each offers an important improvement to the assessment of outcome within their disease areas. The measures showed good psychometric properties across a range of analyses. All of the measures showed good fit to the Rasch model showing the measures were unidimensional and had good measurement properties. Classical Test Theory (CTT) analyses confirmed low levels of random measurement error. In addition, all scales were able to distinguish between known groups showing evidence of construct validity.

4.4.1 Construct definition

The PRIMUS QoL scales and the LCOPD take the needs-based approach to QoL as their conceptual basis. The needs-based approach to QoL is the most widely implemented approach to QoL and differs importantly from measures of Health Related Quality of Life (HRQL) (McKenna, 2011). HRQL measures are usually index based outcomes with varying numbers of domains representing different aspects of the impact of the disease on the patient. They predominantly measure symptoms and functional limitations with just a few items measuring the impact of these on the patient's life. In contrast the needs-based approach to QoL conceptualises quality of life as a unidimensional construct.

The needs-based approach to QoL is embedded within the wider body of theoretical work on human motivation (Maslow, 1970; Max-Neef et al, 1991; Kenrick et al, 2010). The approach postulates that individuals are driven by their needs and that fulfilment of their needs provides satisfaction. Quality of life is high when more needs are fulfilled and lower when they are less able to meet their needs. Chronic diseases impact on the patients' QoL by limiting their ability to meet their needs. Each disease impacts on patients' ability to meet their needs in different ways.

The purpose of the patient interviews in each disease is to identify how that particular illness limits patients' ability to meet their needs. Information from the interviews is used to develop a conceptual framework for the PRO. This includes all of the themes identified and shows how these relate to the patients' needs. During the psychometric evaluation stage care is taken to ensure that the content of the PRO is refined without compromising its conceptual basis.

The conceptual basis of the symptoms and activity limitations scales of the PRIMUS are based on the World Health Organization's (WHO) classification of impairments (physiological and anatomical) and activity limitations (capacity and performance) (World Health Organisation, 1980; 1999). This classification system provides a detailed description of different types of impairments and functional limitations. It also describes the relationship between these and other outcomes. The classification system gives a level of detail that allows very specific hypotheses to be generated from its content, meaning it fulfils much of the criteria necessary for a theory (Popper, 1959). WHO defines impairments as loss or abnormality of psychological, physiological or anatomical structure or function, which represents disturbances at the level of the organ. Activity limitation (functioning) is defined as any restriction or lack of ability to perform an activity in the manner or within the range considered normal for a human being.

The U-FIS was based on the original Fatigue Impact Scale (Fisk et al, 1994). The FIS is a 40-item questionnaire consisting of three sub-scales assessing the impact of perceived fatigue on; cognitive functioning (10 items), physical functioning (10 items) and psychosocial functioning (20 items). Due to the strong inter-relations between the different causes of fatigue and the problems in identifying sub-domains with any clinical validity the U-FIS aimed to capture the overall impact of fatigue as a unidimensional construct (Penner et al, 2007). The conceptual basis of the final measure was again based on the WHO classification of impairments (physiological and anatomical) and activity limitations (capacity and performance) (World Health Organisation, 1980; 1999). The U-FIS is a summary measure of the patient-perceived impact of MS-related fatigue on functional capacity. Consequently, it measures the construct of patient-perceived fatigue-related functional impairment. The scale is not designed as an objective or clinical measure of fatigue symptoms.

Although all scales have specified conceptual foundations they are also somewhat limited in their definitions. As discussed in chapter 2, recent research has attempted to identify the underlying measurement mechanism of the construct (Stenner et al, 2013). By providing this it would be possible to explain 'how' the measures work. This would mean it would be possible to identify the characteristics of the items that make them represent different levels of the construct.

4.4.2 Psychometric methods

Although all studies provided evidence of construct validity, additional assessments of validity would be also desirable. For example, it was not possible to show evidence of the responsiveness of the scales within the studies. Further studies to assess responsiveness are necessary.

The Rasch model was used for item reduction in all of the studies. All of the final scales showed good fit to the model. This means that it is possible for the measures to achieve fundamental measurement. This provides a major advance in measurement in these disease areas.

Despite this some limitations in the methods are worthy of discussion. Methods for assessing fit to the Rasch model evolve over time as knowledge about model requirements is developed. For example, one of the requirements of the Rasch model is local independence of the items. This means items should not be too closely related such that the response to one item has too strong an influence over answers to another item (Tennant and Conaghan, 2007; Baghaei, 2008). It is assessed by identifying items with high residual correlations (after the underlying trait is conditioned out). Correlations should be close to 0. However, there is no clear criterion for identifying high residual correlations. When the scales above were developed a criterion of identifying residual correlations of > 0.3 was used (Tennant and Conaghan, 2007). However, more recently this has been challenged and researchers have argued that this is too lenient and that any correlation value > 0.2 above the average correlation should be used (Christensen et al, 2013). By changing the criteria in this way more evidence of local dependence may have been observed than originally identified. This could have affected the items selected for the measure. In order to overcome this problem scales would need to be continuously re-assessed as new knowledge is gathered. There are obvious practical limitations in how frequently this could be done.

The sample sizes available for Rasch analysis in each study were: PRIMUS, n=135; U-FIS, n = 135; LCOPD UK, n = 162, US, n = 145. As with other statistical analyses, small sample sizes produce less precise and robust estimates and less powerful fit analyses (Linacre, 1994). The sample sizes in each of the studies were large enough to provide 99% confidence that item locations were stable within 0.5 logits (Linacre, 1994). Although this provided a good level of accuracy for the analyses it may be argued that larger samples providing a higher degree of accuracy are necessary for scale development. A sample of size of two hundred and fifty patients or more would give a ‘high stakes’ level of accuracy (Linacre, 1994). Unfortunately, accessing this number of patients in health research is often challenging as many diseases affect only a small proportion of the population.

Finally, only one patient sample was used for the psychometric analyses in each study. As the psychometric study was used to reduce the item pool it means that several items were removed during the analyses. Ideally, a second psychometric study should have been conducted in which the functioning of the scale was assessed using the final item set only. This would confirm that the measure works appropriately without the additional items. Due to the challenge of recruiting such large numbers of patients it was not possible to do this. Further studies should be used to confirm the measurement properties of the scales using the final set of items only.

4.5 Chapter summary

The three articles included in this chapter each describe the development of new disease-specific PROs. The new measures provide important advancements in outcome measurement in their relevant disease areas. The studies showcase a well-developed standardized methodological approach to PRO development. All of the scales had a clear conceptual foundation and showed good fit to the Rasch model.

Some limitations of the research were identified. Further work is necessary to identify the measurement mechanism of the constructs. This will bring a higher degree of clarity to the measures and explain clearly *how* they work. In addition, further Rasch based analyses with larger samples and using more recent analytical methods is desirable.

Chapter 5: Application of new patient-reported outcomes in international research

5.1 Introduction and articles

This chapter discusses three studies relating to the application of PROs in international research studies. Important considerations include the cross-cultural suitability of the construct being measured, the language adaptation of the questionnaire content, cross-cultural differences in the functioning of the scale and how to interpret the scores generated from the PRO.

5.1.1 Article 1: Adapting the Asthma Life Impact Scale (ALIS) for use in Southern European (Italian) and Eastern European (Russian) cultures

Twiss J, McKenna SP, Crawford SR, Tammaru M, Oprandi NC. Adapting the Asthma Life Impact Scale (ALIS) for use in Southern European (Italian) and Eastern European (Russian) cultures. J Med Econ. 2011;14(6):729-38. doi:10.3111/13696998.2011.615356.

BACKGROUND: The Asthma Life Impact Scale (ALIS) is a disease-specific measure used to assess the quality-of-life of people with asthma. It was developed in the UK and US and has proven to be acceptable to patients, to have good psychometric properties, and to be unidimensional.

OBJECTIVE: This paper reports on the adaptation and validation of the ALIS for use in representative Southern European (Italian) and Eastern European (Russian) languages.

METHODS: The ALIS was translated for both cultures using the dual-panel process. The newly translated versions were then tested with asthma patients to ensure face and content validity. Psychometric properties of the new language versions were assessed via a test?re-test postal survey conducted in both countries.

RESULTS: Linguistic nuances were easily resolved during the translation process for both language adaptations. Cognitive debriefing interviews (Russia n=9, male=11.1%, age mean (SD)=55.4 (13.2); Italy n=15, male=66.7%, age mean (SD)=63.5 (11.2)) indicated that the ALIS was easy to read and acceptable to patients. Psychometric testing was conducted on the data (Russia n=61, age mean (SD)=40.7 (15.4); Italy n=71, male=42.6%, age mean (SD)=49.5 (14.1)). The results showed that the new versions of the ALIS were consistent (Russian and Italian Cronbach's alpha=0.92) and reproducible (Russian test-re-test=0.86; Italian test-re-test=0.94). The Italian adaptation showed the expected correlations with the NHP and the Russian adaptation showed strong correlations with the CASIS and CAFS and weak-to-moderate correlations with %FEV1 and %PEF. In both adaptations the ALIS was able to distinguish between participants based on self-reported general health, self-reported severity, and whether or not they were hospitalized in the previous week.

LIMITATIONS: It is possible that some cultural or language differences still exist between the different language versions. Further research should be undertaken to determine responsiveness. Further studies designed to determine the clinical validity of the Italian ALIS would be valuable.

The article can accessed via:
http://www.tandfonline.com/doi/abs/10.3111/13696998.2011.615356?journalCode=ijme20

5.1.2 Article 2: International development of the patient-reported outcome indices for multiple sclerosis (PRIMUS)

McKenna SP, Doward LC, Twiss J, Hagell P, Oprandi NC, Fisk J, Grand'Maison F, Bhan V, Arbizu T, Brassat D, Kohlmann T, Meads DM, Eckert BJ. International development of the patient-reported outcome indices for multiple sclerosis (PRIMUS). Value Health. 2010 Dec;13(8):946-51. doi:10.1111/j.1524-4733.2010.00767.x.

BACKGROUND: The Patient-Reported Indices for Multiple Sclerosis (PRIMUS) comprises a suite of three scales for assessing symptoms, activity limitations, and quality of life in multiple sclerosis (MS). It was developed in the UK and has been shown to have excellent psychometric properties. This study describes the adaptation of eight language versions for Canadian English, Canadian French, French, German, Italian, Spanish, Swedish, and US English.

METHODS: The PRIMUS was translated using the dual-panel process. Cognitive debriefing interviews conducted with MS patients assessed face and content validity. Psychometric and scaling properties were assessed via a two-administration postal survey conducted in each country involving the PRIMUS, the Nottingham Health Profile (NHP), the Unidimensional Fatigue Impact Scale (U-FIS), and demographic questions.

RESULTS: Cognitive debriefing interviews demonstrated the acceptability of the new language versions. Analysis of survey data showed that the new language versions of the three PRIMUS scales were unidimensional (as indicated by fit to the Rasch model) and that they had good internal consistency and reproducibility. PRIMUS scale scores correlated as expected with those on the NHP and the U-FIS. The scales in all countries were able to discriminate between groups of patients on the basis of their self-reported MS severity, general health, and employment status.

CONCLUSIONS: The PRIMUS was successfully adapted into eight new languages. Most of the tests showed the PRIMUS to have good unidimensionality and to have good internal consistency, reproducibility, and construct validity. The measure is now available for use in clinical studies and trials involving these countries and the UK. Further work is required to assess the measure's responsiveness.

The article can be accessed via: http://onlinelibrary.wiley.com/doi/10.1111/j.1524-4733.2010.00767.x/full

5.1.3 Article 3: Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS)

Twiss *et al. Health and Quality of Life Outcomes* 2010, **8**:117
http://www.hqlo.com/content/8/1/117

RESEARCH Open Access

Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS)

James Twiss[1*], Lynda C Doward[1], Stephen P McKenna[1], Benjamin Eckert[2]

Abstract

Background: The PRIMUS is a Multiple Sclerosis (MS)-specific suite of outcome measures including assessments of QoL (PRIMUS QoL, scored 0-22) and activity limitations (PRIMUS Activities, scored 0-30). The U-FIS is a measure of fatigue impact (scored 0-66). These measures have been fully validated previously using an MS sample with mixed diagnoses. The aim of the present study was to validate the measures further in a specifically Relapse Remitting MS (RRMS) sample and to provide preliminary evidence of the responder definitions (RD; also known as minimal important difference) for these instruments.

Methods: Data were derived from a multi-country efficacy trial of MS patients with assessments at baseline and 12 months. Baseline data were used to assess the internal reliability and validity of the measures. Both anchor-based and distribution-based approaches were employed for estimating RD. Anchor-based estimates were based on published RD values for the EQ-5D and were assessed for those improving and deteriorating separately. Distribution-based estimates were based on standard error of measurement (SEM), change score equivalent to 0.30, and change score equivalent to 0.50, effect sizes (ES).

Results: The sample included 911 RRMS patients (67.3% female, age mean (SD) 36.2 (8.4) years, duration of MS mean (SD) 4.8 (5.2) years). Results showed that the PRIMUS and U-FIS had good internal consistency. Appropriate correlations were observed with comparator instruments and both measures were able to distinguish between participants based on Expanded Disability Status Scale scores and time since diagnosis. The anchor-based and distribution-based RD estimates were: PRIMUS Activities range = 1.2-2.3, PRIMUS QoL range = 1.0-2.2, and U-FIS range = 2.4-7.0.

Conclusions: The results show that the PRIMUS and U-FIS are valid instruments for use with RRMS patients. The analyses provide preliminary information on how to interpret scores on the scales. These data will be useful for assessing treatment efficacy and for powering clinical studies.

Trial Reference Number: ClinicalTrials.gov Identifier NCT00340834.

Background

Multiple sclerosis (MS) is a chronic, autoimmune and neurodegenerative disorder of the central nervous system (CNS) characterized by inflammation, demyelination and neuronal loss. MS represents the leading cause of non-traumatic neurologic disability in young and middle-aged adults, affecting an estimated 2.5 million individuals worldwide [1]. About 85% of patients begin with the Relapse Remitting form of MS (RRMS) which is characterised by episodes of symptoms followed by resolution, at least partly, within days to months [2,3]. The long term clinical effects of MS often lead to serious disability. Symptoms of MS are wide ranging and can include weakness of the limbs (particularly the legs), fatigue, unsteadiness, difficulty with bladder control, visual changes due to the involvement of the optic nerve, vertigo, facial numbness or weakness or double vision [4]. In addition, depression occurs in about a quarter of patients [5]. Unsurprisingly, the disease can have major detrimental effects on a patient's QoL [3,6,7].

* Correspondence: JTwiss@Galen-Research.com
[1]Galen Research Ltd, Manchester, UK
Full list of author information is available at the end of the article

Measuring the wide ranging effects of MS is important for developing understanding and treatment of this disease. The Patient Reported Indices for Multiple Sclerosis (PRIMUS) was developed to capture the overall impact of MS from the patient's perspective [8]. This instrument consists of three distinct scales specific to MS; symptoms, activity limitations and quality of life (QoL), each designed to be used in combination or as a standalone measure. Scale content was generated directly from MS patients and, consequently closely represents patients' experience of MS. As fatigue is present in about three quarters of patients [9] the Unidimensional Fatigue Impact scale (U-FIS) [10] was developed in parallel with the PRIMUS scales to provide an index of the impact of fatigue associated with MS. The PRIMUS and U-FIS scales were developed and validated in patients representing the most common MS sub-types; RRMS, Secondary Progressive MS and Primary Progressive MS [8,10]. Data from a large 12 month efficacy trial were made available to evaluate the validity of the instruments further specifically for RRMS. These data also provided an opportunity to investigate how to interpret scores for the PRIMUS and U-FIS.

One of the most commonly used approaches for investigating how to interpret scores on Patient Reported Outcome (PRO) scales has been through the calculation of a minimum score that can be considered to be clinically meaningful. This score can then be used to help interpret treatment response during therapeutic trials. Calculation of this score has been referred to as the Minimal Important Difference (MID) [11], meaningful change [12] and minimal clinically significant difference [13]. More recently the term Responder Definition (RD) has replaced previous terminology [14].

No single method for estimating the RD is widely accepted. Approaches can be classified broadly into anchor-based and distribution-based approaches. Anchor-based approaches involve relating change scores on the PRO to change in a factor of known importance. These methods usually involve using other PROs, [11,15,16] clinical variables [17,18] or patient global rating of change questions [12,19,20] as an anchor. Each approach has strengths and limitations. Other comparator instruments can only be used when the instruments are suitably related to the testing instrument and cover issues important and relevant to the patient [21]. Some authors have suggested that a correlation of 0.5 is necessary between the anchor and main instrument in order to ensure adequate relatedness [15,16]. In these cases it is also useful if previous research has investigated the RD of the comparator instrument. Clinical variables can provide useful markers for interpreting scores on PROs but they do not provide minimal important difference estimates per se. These are most useful when other information for estimating RD is unavailable. Global Rating of Change (GRC) questions generally have multiple Likert type response options ranging from 'very much worse' to 'very much better'. Change scores for those individuals responding 'a little' or 'moderately' improved are used to estimate the RD. Although global rating of change questions are easy to administer the reliability of such methods is questionable. Doubt exists about whether patients can recall their health over periods of time and it is unknown whether patients respond primarily in relation to their current health rather than their change in health [22]. It has also been argued that estimation of RD should not be based on GRC items alone [21].

Distribution-based approaches assess the distribution of scores on the PRO and attempt to identify a score that may be considered important above the 'statistical noise' of the measure. Various distribution-based approaches have been suggested including effect size [23], half a standard deviation [24], the standard error of measurement (SEM) [25] and the standard response mean (SRM) [26]. These different approaches usually produce different magnitudes of RD. Furthermore, distribution-based estimates can sometimes differ considerably from those obtained using anchor-based methods [27].

No previous study has attempted to determine the RD of the PRIMUS and U-FIS. The aim of the present study was twofold. First, to provide further evidence of the validity of the PRIMUS and U-FIS in a RRMS sample. Secondly, to investigate the RD of the PRIMUS and U-FIS scales.

Methods

Patients

Analyses were based on data collected in a 12-month, randomized, multicenter, double-blind, efficacy trial where patients were randomized to receive a fixed dose of either FTY720 0.5 mg/day orally, FTY720 1.25 mg/day orally or interferon beta-1a 30 μg/week. The trial included 1292 RRMS patients at 172 centers in 18 countries. PRIMUS and U-FIS data were only available for countries where the questionnaires had been previously formally adapted and validated [8,28,10,29]. Data were available for 911 patients from the following 8 countries; Canada (French and English), France, Germany, Italy, Spain, United Kingdom, United States and Australia.

The participants were aged 18 to 55 years, with active MS (defined as one relapse during the previous year or two relapses during the previous 2 years), Expanded Disability Status Scale (EDSS) score of between 0 and 5.5 and neurologically stable for at least 30 days prior to randomization.

Measures

The PRIMUS consists of three independent scales; symptoms, activity limitations and QoL designed to be used as standalone measures or in combination [8,28]. For the present study data were available for the QoL and activity limitation scales. The QoL scale contains 22-items in the form of simple statements accompanied by dichotomous response options. Items are summed in each scale to yield a total score ranging from 0 to 22. High scores indicate worse QoL. The activity limitations scale contains 15-items describing specific physical tasks. Respondents rate the degree to which they are able to perform the tasks on a three point scale. Again, items are summed to give a total score that can range from 0 to 30. High scores are indicative of greater activity limitation. Both scales have been shown to be unidimensional and to have good reproducibility and validity in a number of languages [28].

The U-FIS has 22-items measuring the impact of fatigue [10,29]. For each item, individuals rate the degree to which they have been affected by fatigue during the previous week on a scale ranging from 'Never' (scored 0) to 'All the time' (scored 3). Item scores are summed to give a total score that can range from 0 to 66. The U-FIS is unidimensional and has been shown to have good reproducibility and validity in several languages [29]. The PRIMUS and U-FIS are available at http://www.galen-research.com.

The Expanded Disability Status Scale (EDSS) is a global scale developed to evaluate disability due to neurologic limitations in people with MS [30]. It has 20 available levels that describe progressive disability ranging from 0 (normal) to 10 (death due to MS) rising in 0.5 units. Patients are clinically assessed and assigned scores in eight functional systems that are scored from 0-5 or 0-6. Higher scores represent greater system impact. The eight functional systems are; pyramidal, cerebellar, brainstem, sensory, bowel and bladder, visual and cerebral/mental functions. EDSS scores are generated from the system functions scores and other information collected during the clinical examination.

The Multiple Sclerosis Functional composite (MSFC) is a clinical measure of physical and cognitive functioning in MS patients [31]. It assesses leg function/ambulation, arm/hand function and cognitive function. These three scales are also added together to give a composite measure of functioning. The leg function/ambulation measure is based on the average of two timed 25-foot walk tests. The arm/hand function measure involves four 9-hole peg tests. The cognitive function measure is the Paced Auditory Serial Addition Test (PASAT) that assesses auditory processing speed and working memory [32]. The three separate scale scores are converted into z-scores before being added together to form a composite score.

The EQ-5D is a generic health outcome assessment [33]. It consists of 5 items: Mobility, Self-care, Usual activities, Pain/Discomfort and Anxiety/depression, each with 3 levels (no problems, moderate problems, extreme problems). A health utility value is derived for each patient based on their combination of responses to the five items. The score is on a continuum from 1 (best possible health) to 0 (death) with some health states being valued worse than death (< 0). Research has suggested that the RD of the EQ-5D is 0.074 [34].

Statistical analysis

Reliability and Validity

The distributional properties of the PRIMUS and U-FIS were explored through descriptive statistics (mean, standard deviation, median and inter-quartile range [IQR]) and floor and ceiling effects (percentage of patients scoring the minimum and maximum possible scores, respectively). Internal consistency (degree of relatedness of items) was assessed using Cronbach's alpha. A correlation of 0.70 is accepted as indicating adequate consistency [35]. Convergent and discriminant validity were evaluated by assessing the level of association (Spearman rank correlations) between scores on the PRIMUS and U-FIS scales and those on the EQ-5D, EDSS and the MSFC subscales and composite score. Known groups validity was assessed by examining the PRIMUS and U-FIS scores of respondents who differed according to their baseline EDSS group and duration of MS. EDSS group was defined in the following way; EDSS (0 - 1.5), EDSS (2 - 2.5), EDSS (3 - 3.5), EDSS (4-5.5). Non-parametric tests for independent samples (Mann-Whitney U Test for two groups and Kruskal-Wallis one-way analysis of variance for three or more groups) were employed. Psychometric testing was performed using the SPSS 17.0 statistical package.

Responder Definition Analysis

The RDs for the PRIMUS and U-FIS were estimated using a combination of anchor-based and distribution-based methods. Anchor-based analyses were conducted by comparing scores on the PRIMUS and U-FIS with published RD values for the EQ-5D [34]. The anchor approach assessed change scores for the PRIMUS and U-FIS for individuals who improved or deteriorated by 0.074-0.111 on the EQ-5D (1-1.5 times the RD of the EQ-5D).

The distributional methods included the assessment of effect size, half a standard deviation and standard error of measurement. The effect size (ES) statistic is based on the ratio of difference between a target measure's mean at baseline and at follow-up (related to the standard deviation of the baseline scores). The group change ES is calculated as follows:

$$ES = \frac{(m_2 - m_1)}{s_1}$$

Where m_1 is the group mean at baseline, m_2 is the group mean at follow-up and s_1 is the group standard deviation at baseline. Cohen devised ES thresholds for assessing the magnitude of group change that are widely accepted [23]. These are 0.2 for a small group change, 0.5 for a moderate group change and 0.8 for a large group change. Estimates of change scores needed to produce different effect sizes can be calculated using baseline standard deviations. Half a standard deviation (equivalent to half the baseline standard deviation) is commonly found to be close in value to published RD values [24]. Change scores required to produce effect sizes of 0.3, and 0.5 were calculated.

The SEM has also been posited as a surrogate for the RD [25]. It has been described as the standard error in an observed score that obscures the true score [36]. It is estimated as follows:

$$SEM = s_1 \times \left(\sqrt{1 - r}\right)$$

Standard deviation at baseline (s_1) is multiplied by the square root of one minus the internal consistency of the target measure (as assessed by Cronbach's Alpha coefficient (r)). SEM has been used frequently to aid in the interpretation of PRO scores and a change above 1 SEM has been considered to be meaningful [37-40].

Results

Demographic and disease information for the sample is shown in Table 1. The table shows that the sample was relatively mild in terms of MS severity. A majority of patients had EDSS scores between 0 and 2.5 and most reported having had two or fewer relapses in the previous two years.

Questionnaire responses on the PRIMUS, U-FIS and EQ-5D are reported in Table 2. Results showed that over 20% of respondents scored the minimum for the PRIMUS Activity limitations and QoL scale and the maximum for the EQ-5D scale (which indicates good health status). These findings confirm the relatively low baseline disability in the sample. Results showed that there were few signs of ceiling effects for the PRIMUS or U-FIS scales.

Internal consistency

Cronbach's alpha coefficients for the scales were; PRIMUS Activities 0.88, PRIMUS QoL 0.92, and U-FIS 0.97. As cronbach's alpha coefficients were all above 0.7 this indicated good interrelatedness of items.

Convergent validity

Correlations between questionnaire and physician assessments are shown in Table 3. As anticipated, moderate correlations were found between the PRIMUS

Table 1 Participant details (n = 911)

Sex		
	Male (%)	292 (32.1)
	Female (%)	618 (67.8)
	Missing (%)	1 (0.1)
Age (years)		
	Mean (SD)	36.5 (8.4)
	Median (IQR)	37 (30 - 43)
	Range	18 - 55
	Missing (%)	0
Duration of MS (years)		
	Mean (SD)	4.8 (5.2)
	Median (IQR)	3.2 (0.7 - 7.2)
	Range	0.1 - 32.9
	Missing (%)	9 (1)
Number (%) relapses in the previous 2 years		
	1	268 (29.4)
	2	536 (58.8)
	3	86 (9.4)
	4	18 (2.0)
	Missing (%)	3 (0.3)
EDSS Group (%)		
	0-1.5	400 (44.3)
	2-2.5	262 (29.0)
	3-3.5	135 (15.0)
	4 +	105 (11.6)
	Missing (%)	9 (1)

Table 2 Descriptive scores on patient reported outcome measures

	PRIMUS QoL	PRIMUS Activities	UFIS	EQ-5 D Utility
Baseline				
N	885	883	873	900
Mean (SD)	4.0 (4.3)	3.0 (4.6)	16.8 (13.9)	0.80 (0.19)
Median (IQR)	2.0 (1.0 - 6.0)	2.0 (0 - 4.0)	14.0 (5.0 - 27.0)	0.80 (0.73 - 1)
% scoring Min	21.4	39.8	7.0	0
% scoring Max	0	0.2	0	29.9
12 Months				
n	835	833	825	839
Mean (SD)	3.8 (4.7)	3.2 (4.8)	17.0 (14.8)	0.80 (0.21)
Median (IQR)	2.0 (0 - 6.0)	1.0 (0 - 4.0)	13.0 (4.0 - 27.0)	0.81 (0.73 - 1)
% scoring Min	29.8	41.5	10.4	0
% scoring Max	0.2	0.4	0.2	35.2

Table 3 Convergent validity PRIMUS QoL, PRIMUS Activities and U-FIS at baseline

	PRIMUS QoL	PRIMUS Activities	U-FIS	Timed 25 foot Walk test	9-hole peg test	PASAT	MSFC Total	EDSS
PRIMUS Activities	.62							
U-FIS	.75	.66						
Timed 25 foot Walk test	.20	.32	.22					
9-hole peg test	.20	.31	.22	.31				
PASAT	-.17	-.18	-.18	-.20	-.20			
MSFC Total	-.24	-.33	-.25	-.47	-.72	.71		
EDSS	.35	.65	.38	.27	.34	-.14	-.31	
EQ-5 D Utility	-.58	-.58	-.60	-.20	-.23	.14	.24	-.35

All correlations were significant at the <0.01 level (2 tailed, Spearman Rank correlations)

scales/U-FIS and EQ-5D scales as these assess related but distinct constructs. The PRIMUS scales and the U-FIS correlated strongly with each other. The EDSS showed low to moderate correlations with the PRIMUS scales and with the U-FIS. The PRIMUS QoL scale and the U-FIS showed weak associations with the MSFC scales and composite score. The PRIMUS Activities scale showed slightly stronger associations with the MSFC scales and composite but these still remained lower than expected. It should be noted that the EDSS and the EQ-5D also showed lower than expected correlations with the MSFC composite score and its subscales. In particular, all scales correlated weakly with the MSFC PASAT scores.

Known group validity

Results of the known group validity assessments for the PRIMUS and U-FIS sales are shown in Table 4. Each of the scales was able to distinguish between participants based on EDSS group. As expected, individuals with greater disability according to EDSS had significantly higher PRIMUS and U-FIS scores. The PRIMUS scales and U-FIS were also able to distinguish between participants based on their duration of MS. As anticipated, individuals who had experienced MS for longer had significantly higher scores on the scales. The PRIMUS scales and U-FIS were also able to distinguish between individuals based on the number of relapses they had experienced in the previous two years. Significant differences in PRIMUS activity limitations and U-FIS scores were found between groups split by number of relapses in the previous two years. Individuals with more relapses obtained higher scores. There was a similar, but not statistically significant, finding for QoL scores. However, both the PRIMUS QoL and U-FIS scales showed statistically significant differences between patients who reported two relapses compared with those who reported three or more.

Table 4 Known Group Validity at baseline

		PRIMUS QoL		PRIMUS Activities		UFIS
	n	Mean (SD)	n	Mean (SD)	n	Mean (SD)
EDSS Group						
0-1.5	391	2.7 (3.5)	393	1.6 (3.5)	381	11.7 (11.0)
2-2.5	255	3.8 (4.0)	253	2.7 (3.8)	252	17.6 (13.7)
3-3.5	130	5.3 (4.6)	129	4.5 (5.4)	129	22.2 (14.4)
4-5.5	102	7.4 (5.2)	99	7.7 (5.5)	102	27.1 (14.8)
P		< 0.01		< 0.01		< 0.01
Number of relapses in previous 2 years						
1	259	3.8 (4.3)	260	2.2 (3.4)	262	16.1 (13.8)
2	522	3.8 (4.1)	519	3.1 (4.7)	508	16.2 (13.4)
3+	101	5.1 (5.3)	101	4.7 (6.3)	100	22.1 (15.4)
P		0.084		< 0.01		< 0.01
Median MS duration group						
Below median (3.2)	439	3.6 (4.2)	435	2.3 (4.1)	435	14.5 (13.3)
Above median (3.2)	439	4.3 (4.4)	439	3.8 (5.0)	429	19.1 (14.1)
P		< 0.01		< 0.01		< 0.01

Non-parametric tests were conducted (Mann-Whitney U Test for two groups and Kruskal-Wallis one-way analysis of variance for three or more groups)

Responder definition analysis

The anchor-based estimates for the RD for those improving and deteriorating are shown in Table 5. The results showed that for the PRIMUS Activities and QoL scales the RD estimates were similar for patients who improved or deteriorated. There was a more pronounced difference in RD estimates between patients who improved or deteriorated according to the U-FIS. Note that scores for no change in EQ-5D provided the following change scores; -0.2 (n = 331) for Activity limitations, 0.3 (n = 331) for QoL and 0.0 (n = 325) for U-FIS.

Values for the distribution-based approaches (SEM and ES) are also shown in Table 5. The distribution-based estimates provided similar values to the anchor-based estimates.

The final ranges in RD values for each scale were PRIMUS QoL 1.0-2.2, Activities 1.2-2.3 and U-FIS 2.4-7.0.

Discussion

The results of this study support the use of the PRIMUS and U-FIS with Relapse Remitting MS samples. Questionnaire descriptive statistics confirmed the mild severity of the sample demonstrated by the clinical data. Internal consistency was above 0.70 for the PRIMUS and U-FIS scales indicating that items in the scales were sufficiently related. Convergent and divergent validity showed that the PRIMUS and U-FIS scales had the expected patterns of association with the comparator measures. Scores on the PRIMUS and U-FIS scales were also related to each other in the same way as was found in previous research involving a wider range of types of MS [8,10]. Associations between the PRIMUS and U-FIS and the MSFC subscales and composite score were weaker than expected. However, associations between the MSFC, EDSS and EQ-5D were also weaker than expected suggesting that further investigation of the relation between the MSFC and other clinical outcome measures is needed [41-44].

Known groups validity results showed that the PRIMUS scales and the U-FIS were able to distinguish between participants based on their EDSS level and duration of illness. The PRIMUS and U-FIS scales were also able to distinguish between participants based on the number of relapses they had experienced in the previous two years, although, this difference was not statistically significant for the PRIMUS QoL scale. However, it may be more appropriate to measure relapse frequency yearly or 6 monthly to provide more accurate information.

The anchor estimates produced preliminary evidence of the RDs for the PRIMUS and U-FIS. Encouragingly, the scores obtained for the anchor-based estimates were similar in value to those obtained from the distribution-based estimates. Previous research has suggested that there may be differences in RD values depending on whether individuals improve or deteriorate [45-47]. In the present study there was no bi-directional difference in anchor-based RD values for individuals who improved or deteriorated for the PRIMUS Activities and QoL scales. However, there was a bi-directional difference for the U-FIS; individuals who improved had an RD of 6.5 compared with 4.7 for those who deteriorated. Despite this difference both the improving and deteriorating anchor values for the U-FIS were within the range of the distribution-based estimates. It is unclear whether there are true differences in the RD values for individuals with improving or deteriorating scores on the U-FIS. Further research is needed to investigate this issue.

The final range in values for each scale can be used to provide preliminary guidance when interpreting changes in scores on the measures and to aid calculation of sample sizes needed for clinical studies. Future research is needed to determine whether the RD estimates remain constant in more severe samples and with different types of MS. Previous researchers have highlighted the possibility that the RD may vary as a function of severity [13,21]. For example, it is possible that individuals with

Table 5 Responder definition estimates

	PRIMUS Activities		PRIMUS QoL		U-FIS	
Responder Definition Estimates	**n**	**Mean change score**	**n**	**Mean change score**	**n**	**Mean change score**
Anchor-based						
Equivalent to reported EQ-5 D RD improvement	16	-2.3	17	-1.0	17	-6.5
Equivalent to reported EQ-5 D RD deterioration	15	1.8	15	1.0	15	4.7
Distribution-based						
SEM	814	1.2	812	1.5	815	2.4
Change score equivalent to 0.30 ES	814	1.4	812	1.3	815	4.2
Change score equivalent to 0.50 ES	814	2.3	812	2.2	815	7.0

Distribution-based estimates were conducted at baseline

severe forms of Secondary Progressive MS may have higher RDs for the scales. The present study investigated the RDs of the PRIMUS and U-FIS in a fairly mild sample of RM patients and the results can be considered valid for future similar samples.

The study has a number of limitations. As mentioned earlier, the sample included a high proportion of patients at the low end of the MS disability spectrum. However, this is consistent with recent clinical trials of RRMS patients and is likely to be reflected in future RRMS studies where the PRIMUS and UFIS are applied. The present assessments were unable to report on the reproducibility of the PRIMUS and U-FIS scales in this sample. However, previous research, including a large proportion of RRMS patients, indicated that the scales had excellent reproducibility [8,10,28,29]. Anchor-based estimates of RD were based on the published RD value for the EQ-5D. Although this provided a useful tool for the present study there are other potential anchors that could be used such as a global question on change in overall health. Finally, as there was little change in patient condition during the trial, relatively few patients could be included in the RD anchor analysis.

Conclusions

The PRIMUS and U-FIS have been shown to be reliable and valid instruments for the assessment of outcome in RRMS patients. RD estimates are between 1.2-2.3 for the PRIMUS Activity scale, 1.0-2.2 for the QoL scale and 2.4-7.0 for the U-FIS. These estimates are important to help interpretation of change scores and to assist in determining sample sizes necessary for future clinical studies.

Abbreviations

MID: minimal clinically significant difference; MS: multiple sclerosis; QoL: quality of life; PRO: patient reported outcome; RD: responder definition; RRMS: Relapse Remitting Multiple Sclerosis.

Acknowledgements

This study was funded by Novartis Pharmaceuticals. We are grateful to all participants who completed the questionnaires.

Author details

[1]Galen Research Ltd, Manchester, UK. [2]Global Health Economics and Outcomes Research, Novartis Pharmaceuticals, Basel, Switzerland.

Authors' contributions

JT was involved with the design of the study, analysis and interpretation of data and drafting of the manuscript. LCD was involved in the conception and design of the study, interpretation of data and contributed to the manuscript. SPM was involved with the design of study, interpretation of the data and contributed to the manuscript. BE was involved with the design of the study, acquisition of data and reviewed and contributed to the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 11 January 2010 Accepted: 11 October 2010
Published: 11 October 2010

References

1. Multiple Sclerosis International Federation (MSIF):[http://www.msif.org/en/about_ms/]. [accessed 02.12.09]. About MS.
2. Vollmer T: **The natural history of relapses in multiple sclerosis.** *J Neurol Sci* 2007, **256(Suppl 1)**:5-13.
3. Putzki N, Fischer J, Gottwald K, Reifschneider G, Ries S, Siever A, Hoffmann F, Kafferlein W, Kausch U, Liedtke M, Kirchmeier J, Gmund S, Richter A, Schicklmaier P, Niemczyk G, Wernsdorfer C, Hartung HP, for the "Mensch im Mittelpunkt" Study Group: **Quality of Life in 1000 patients with early relapsing-remitting multiple sclerosis.** *Eur J Neurol* 2009, **16**:713-20.
4. Murray JT: *Multiple Sclerosis, the History of a Disease* New York: Demos Medical Publishing 2005.
5. Patten SB, Williams JVA, Barbui C, Metz LM: **Major depression in multiple sclerosis a population based perspective.** *Neurology* 2003, **61**:1524-27.
6. Montel SR, Bungener C: **Coping and quality of life in one hundred and thirty five subjects with multiple sclerosis.** *Mult Scler* 2006, **13**:393-401.
7. Ziemssen T: **Multiple Sclerosis beyond EDSS: depression and fatigue.** *J Neurol Sci* 2009, **277(Suppl 1)**:37-41.
8. Doward LC, McKenna SP, Meads DM, Twiss J, Eckert BJ: **The Development of Patient Reported Outcome Indices for Multiple Sclerosis (PRIMUS).** *Mult Scler* 2009, **15(9)**:1092-1102.
9. Lerdal A, Celius EG, Krupp L, Dahl AA: **A prospective study of patterns of fatigue in multiple sclerosis.** *Eur J Neurol* 2007, **14**:1338-43.
10. Meads D, Doward L, McKenna S, Fisk J, Twiss J, Eckert B: **The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS).** *Mult Scler* 2009, **15**:1228-1238.
11. Pickard SA, Neary MP, Cella D: **Estimation of minimally important differences in EQ-5 D utility and VAS scores in cancer.** *Health Qual Life Outcomes* 2007, **5**:70.
12. Crosby RD, Kolotkin RL, Williams GR: **An integrated method to determine meaningful changes in health-related Quality of Life.** *J Clin Epidemiol* 2004, **57**:1153-1160.
13. Hajiro T, Nishimaru K: **Minimal clinically significant difference in health status: the thorny path of health status measures?** *Eur Respir J* 2002, **19**:390-391.
14. U.S. Department of Health and Human Services Food and Drug Administration Guidance for Industry: **Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.** U.S. FDA; Clinical/Medical 2009 [http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf]. Accessed 9th December 2009.
15. Puhan MA, Frey M, Büchi S, Schünemann HJ: **The minimal important differences of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease.** *Health Qual Life Outcomes* 2008, **6**:46.
16. Schunemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH: **Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction.** *J Clin Epidemiol* 2003, **56(12)**:1170-1176.
17. Santanello NC, Zhang J, Seidenberg B, Reiss TF, Barber BL: **What are minimal important changes for asthma measures in a clinical trial?** *Eur Respir J* 1999, **14**:23-27.
18. Jones PW: **Interpreting thresholds for a clinically significant change in health status in asthma and COPD.** *Eur Respir J* 2002, **19**:398-404.
19. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffith AM, Critch JN, Guyatt GH: **Using the entire cohort in the receiver operating characteristic analysis maximises the precision of the minimal important difference.** *J Clin Epidemiol* 2009, **62**:374-379.
20. Stargardt T, Gonder-Frederick L, Krobot KJ, Alexander CM: **Fear of Hypoglycaemia: defining a minimum clinically important difference in patients with type 2 diabetes.** *Health Qual Life Outcomes* 2009, **7**:91.
21. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR: **Methods to explain the clinical significance of health status measures.** *Mayo Clinic proceedings* 2002, **77(4)**:371-383.
22. Norman GR, Stratford P, Regehr G: **Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach.** *J Clin Epidemiol* 1997, **50**:869-879.

23. Cohen J: *Statistical Power Analysis for the Behavioural Sciences* New York: Academic Press 1977.
24. Norman GR, Sloan JA, Wyrwich KW: **Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation.** *Med Care* 2003, **41**:582-92, Review.
25. Wyrwich KW: **Minimal important difference thresholds and the standard error of measurement: is there a connection?** *J Biopharm Stat* 2004, **14**:97-110.
26. Beaton DE, Hogg-Johnson S, Bombadier C: **Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders.** *J Clin Epidemiol* 1997, **50**:79-93.
27. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffith AM, Critch JN, Guyatt GH: **The minimal detectable change cannot reliably replace the minimal important difference.** *J Clin Epidemiol* 2010, **63**:28-36.
28. McKenna SP, Doward LC, Twiss J, Hagell P, Oprandi NC, Fisk J, Grand'Maison F, Bhan V, Arbizu T, Brassat D, Kohlmann T, Meads DM, Eckert BJ: **International Development of the Patient-Reported Outcome Indices for Multiple Sclerosis (PRIMUS).** *Value Health* 2010.
29. Doward LC, Meads DM, Fisk J, Twiss J, Hagell P, Oprandi N, Goodman J, Grand'Maison F, Bhan V, Gonzalez B, Txomin A, Kohlmann T, Brassat D, Eckert BJ, McKenna SP: **International development of the Unidimensional Fatigue Impact Scale (U-FIS).** *Value Health* 2010, **13(4)**:463-468.
30. Kurtzke JF: **Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS).** *Neurology* 1983, **33**:1444-52.
31. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, Syndulko K, Weinshenker BG, Antel JP, Confavreux C, Ellison GW, Lublin F, Miller AE, Rao SM, Reingold S, Thompson A, Willoughby E: **Development of a multiple sclerosis functional composite as a clinical trial outcome measure.** *Brain* 1999, **122(Pt 5)**:871-82.
32. Gronwall DM: **Paced Auditory Serial-Addition Task: a measure of recovery from concussion.** *Percept Mot Skills* 1977, **44**:367-373.
33. EuroQoL Group: **EuroQoL - a new facility for the measurement of health-related quality of life.** *Health Policy* 1990, **16**:199-208.
34. Walters SJ, Brazier JE: **Comparison of the minimally important difference for two health state utility measures: EQ-5 D and SF-6D.** *Qual Life Res* 2005, **14**:1523-1532.
35. Nunnally JC Jr: *Psychometric Theory* New York: McGraw-Hill 1978.
36. Anastasi A, Urbina S: *Psychological Testing* New Jersey: Prentice Hall 1997.
37. Fitzpatrick R, Norquist JM, Jenkinson C: **Distribution-based criteria for change in health-related quality of life in Parkinson's disease.** *J Clin Epidemiol* 2004, **57**:40-44.
38. Wyrwich KW, Nienaber NA, Tierney WM, *et al*: **Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life.** *Med Care* 1999, **37**:469-478.
39. Wyrwich KW, Tierney WM, Wolinsky FD: **Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life.** *J Clin Epidemiol* 1999, **52**:861-873.
40. Wyrwich KW, Tierney WM, Wolinsky FD: **Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire.** *Qual Life Res* 2002, **11**:1-7.
41. Alvarez-Lafuente R, Garcia-Montojo M, De Las Heras V, Dominguez-Mozo MI, Bartolome M, Garcia-Martinez A, Arroyo R: **A two-year follow-up study: multiple sclerosis functional composite versus expanded disability status scale.** *Mult Scler* 2009, **15(Suppl 9)**:55-56.
42. Kragt JJ, Thompson AJ, Montalban X, Tintore M, Rio J, Polman CH, Uitdehaag BMJ: **Responsiveness and predictive value of EDSS and MSFC in primary progressive MS.** *Neurology* 2008, **70**:1084-1091.
43. Costelloe L, Hutchinson M: **Is a 20% change in MSFC components clinically meaningful?** *Mult Scler* 2007, **13**:1076.
44. Casanova B, Pascual A, Bernat A, Escutia M, Bosca I, Coret F: **Learning effect on multiple sclerosis functional composite in daily clinical practice [abstract].** *Mult Scler* 2004, **10(Suppl 2)**:118.
45. Cella D, Hahn EA, Dineen K: **Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening.** *Qual Life Res* 2002, **11**:207-221.
46. Kwok T, Pope JE: **Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales.** *J Rheumatol* 2010, **37(5)**:1024-8.
47. Colangelo KJ, Pope JE, Peschken C: **The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36.** *J Rheumatol* 2009, **36(10)**:2231-7.

doi:10.1186/1477-7525-8-117
Cite this article as: Twiss *et al.*: Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS). *Health and Quality of Life Outcomes* 2010 **8**:117.

5.2 Description of studies

The first study discussed the adaptation of the Asthma Life Impact Scale (ALIS; Meads et al, 2010) into two new languages – Italian and Russian. The ALIS is a quality of life (QoL) scale specific to patients with Asthma based on the needs-based approach. It was developed using the same development methods discussed in Chapter 4. In this study the measure was translated using a careful adaption procedure into Italian and Russian and then assessed in a psychometric study to test the functioning of the scale. The classical test theory (CTT) psychometric properties of the scale were evaluated (including internal reliability and test-retest reliability) and the construct validity of the new measures assessed. The adaptation procedure was successful and the measure showed good psychometric properties. Internal reliability (Russian and Italian Cronbach's alpha=0.92) and test-retest reliability was acceptable (Russian test–re-test=0.86; Italian test–retest=0.94). In both adaptations the ALIS was able to distinguish between participants based on self-reported general health, self-reported severity of disease and whether or not patients were hospitalized in the previous week.

The second study discussed the international development of the Patient-Reported Outcome Indices for Multiple Sclerosis (PRIMUS; Doward et al, 2009b). The original development was described in detail in Chapter 4 (section 4.3). This study discusses the adaptation of the measure into eight languages; Canadian English, Canadian French, French, German, Italian, Spanish, Swedish and US English. During the adaptation the measure was translated using a standard adaptation procedure and then evaluated in a psychometric study that applied both CTT and Rasch analysis. The measure was translated successfully and showed good fit to the Rasch model in all languages. Adequate levels of internal reliability (Crobnach's alpha >0.70) and test-retest reliability (Spearman rank >0.80) were observed. Validity assessments showed that all language adaptations were able to distinguish between groups based on self-reported MS severity, general health, and employment status.

The third study described the estimation of the minimal important difference (MID) for the PRIMUS and unidimensional fatigue impact scale (U-FIS; Doward et al, 2009b; Meads et al, 2009). The MID is an estimate of the minimum difference that can be considered important from the patients' perspective. In studies with large samples even small differences in PRO scores between groups can show statistical significance (Hochster, 2008). This is because large samples have more statistical power. In these situations it is also important to assess the clinical significance of the differences between the groups. Using the MID as a guide assists in this process. In the study MID values were calculated for the two measures using a range of different methods. The MID study was successful in identifying estimates for the PRIMUS and U-FIS. MID estimates are between 1.2-2.3 for the PRIMUS Activity scale, 1.0-2.2 for the QoL scale and 2.4-7.0 for the U-FIS.

Samples of the UK versions of the PRIMUS, U-FIS and ALIS are provided in Appendices 2, 3 and 5. Copies of the full measures can be obtained from gr@galen-research.com.

5.3 Methodology

Ethics approval was sought and obtained where necessary for the three studies. For the two adaptation studies ethics approval was not required in some of the countries included. Each study required the transfer of data from the countries in which the work was conducted back to the UK. All data was anonymised before the transfer occurred.

It is important that the adaptation of a PRO is based on a thorough adaptation methodology. The language used may contain many nuances and phrases easily understood in the original language that may not be clear to non-native speakers. Consequently, it is inappropriate to produce a new language version of a questionnaire by simply translating the content (literal translation) as it can lead to a poor translation in the target language.

The adaptation methodology used in the ALIS and PRIMUS studies followed the same procedures. Three stages were involved in the adaptations:

1. Translation using dual panel methodology (Swaine-Verdier et al, 2004).

2. Cognitive debriefing interviews.

3. Assessment of psychometric properties.

The dual panel method contrasts with the more frequently applied method of forward and backwards translation. Although this method has become the most frequently used there is no scientific basis for its use compared with other methods (Swaine-Verdier et al, 2004). The dual panel translation method consists of conducting a professional and lay panel in each country. The panels each require between four and seven participants. The professional panel includes bilingual speakers while the lay panel includes monolingual speakers in the target language. Using this method the professional panel works to provide the initial translation in the target language. Emphasis is placed on producing conceptual equivalence for each item rather than a simple word for word translation. As the bilingual panel includes individuals of a higher educational level than the average population the language produced is reviewed by a lay panel more representative of the general population. The lay panel assesses the measure for comprehension and 'naturalness' of language. Research has shown that patients prefer adaptations based on the dual panel method compared with the forward-backward method (Hagell et al, 2010).

Cognitive debriefing interviews were conducted with fifteen relevant patients in each language adaptation to assess the applicability, comprehensiveness and relevance of the translated items. Interviewees completed the questionnaire in the presence of an interviewer who noted any obvious difficulties or hesitation over individual items. Patients were then asked to comment on individual items, instructions and the response format.

After a measure has been translated it is important to evaluate its psychometric properties to ensure the new adaptation works in a similar way to the original. In the ALIS adaptation study the functioning of the two new adaptations was assessed using CTT methods including internal consistency and test-retest reliability. In the PRIMUS study both CTT and Rasch analysis were used to assess the functioning of the adaptations. Rasch analysis allowed the measurement properties of the adaptations to be assessed more thoroughly. In addition, item location ordering was also compared across languages.

To estimate the MID for the PRIMUS and U-FIS two main types of analysis were conducted; anchor-based and distribution-based. Anchor-based approaches attempt to relate change scores on the PRO to change in a factor of known importance. The anchors used in the study included published MID values for the EQ-5D (Walters and Brazier, 2005). Distribution methods attempt to identify a score that may be considered important above the 'statistical noise' of the measure. The distributional methods used in the study were the assessment of effect size, half a standard deviation and standard error of measurement. The final MID values were selected after considering the results produced from all the analyses.

To calculate the MID data from a twelve-month, randomized, multicenter, double-blind, efficacy trial were used. In total nine hundred and eleven patients were available for the analyses from eight countries; Canada (French and English), France, Germany, Italy, Spain, United Kingdom, United States and Australia.

5.4 Evaluation

The adaptions of the ALIS and PRIMUS were successful. The content of the measures were translated with few problems in each language. In addition, good psychometric properties were observed for all language versions in both studies. The PRIMUS adaptation showed how Rasch analysis can be incorporated into the language adaptation process to improve psychometric evaluation. The methods used in both studies go beyond the basic requirements for new adaptations recommended in available guidelines (Wild et al, 2005). In these guidelines, it is recommended that forward-backward translation methods are used and just five patient interviews to assess content validity.

The adaptation of the ALIS into Russian and Italian shows the content of the measure can be easily adapted into two new language groups. Eight new language versions were developed for the PRIMUS. The success of the adaptations shows good evidence for the methodological approach used in the adaptations. The increase in language availability for the new scales means that they are available for future international clinical trials and research studies.

The estimates generated in the MID study are important to help interpret change scores in clinical trials and research studies. In addition, the MID figures also help to determine sample sizes necessary for future clinical studies.

5.4.1 Construct definition

The conceptual basis of the PRIMUS and U-FIS were discussed in Chapter 4. The PRIMUS has three scales; QoL, Activity limitations and symptoms. The U-FIS, PRIMUS Activity limitations and symptoms scales use the World Health Organization's (WHO) classification of impairments (physiological and anatomical) and activity limitations (capacity and performance) (World Health Organisation, 1980; 1999) as their conceptual foundation. The PRIMUS QoL and ALIS are based on the needs-based approach to QoL (McKenna and Doward, 2004).

It is important that the construct being measured is applicable to the target culture. Constructs such as symptoms are less likely to be culturally centric as there should be consistency in the expression of symptoms across cultures. However, QoL may be influenced to a greater extent by the values of a particular culture. The needs-based QoL approach attempts to overcome this by defining QoL based on human needs, which are considered to be universal. Despite this, the way in which needs are satisfied may vary between cultures. This means the content of some PRO items developed in one culture may not be fully relevant in another. This is likely to be of greatest concern when comparing cultures that are very different such as comparing far eastern cultures with the western ones.

The ideal way to ensure the cultural relevance of a PRO is to develop it in several cultures at the same time. This would involve conducting all stages on development in each culture. It is possible for some of the items to be different in each country if different issues arise. As long as there is a core set of items it would be possible to co-calibrate measures in different languages onto the same scale using Rasch analysis (Twiss and McKenna, 2015). Unfortunately, the scope of this work would be very large and has rarely been attempted. Smaller scale studies have been conducted where item generation is performed simultaneously in a small number of countries (McKenna et al, 2003; Whalley et al, 2004).

5.4.2 Psychometric methods

In the ALIS adaptation study psychometric analyses were conducted using CTT methods. These included testing the internal consistency and test-retest reliability of the measures. The results showed that the scale had adequate internal reliability and test-retest reliability. In addition, evidence of construct validity was also provided in the study. The results were similar to those obtained in the original development study. Despite this, CTT methods are limited in the information they can provide regarding measurement equivalence. These methods were applied due to limitations in the sample sizes.

The Rasch model offers a more thorough way of assessing measurement equivalence. In the PRIMUS study overall fit, item level fit and appropriate functioning of the response options was observed for each language separately. This indicates that the scales all worked well in each language. The severity location of the individual items on the underlying scale was also investigated. The mildest and most severe items were found to be the same in each language version.

A more thorough investigation of measurement equivalence would have been provided by a DIF analysis. It was not possible within the scope of the study to assess DIF by language version as this would have required a large level of additional work. This would involve analysing the dataset as a whole and attempting to identify the presence of DIF between languages. If DIF is identified it is necessary to assess the extent to which it influences the calculation of the Rasch estimates. The importance of any identified DIF can be tested using a method outlined by Tenant and Pallant (2007). Estimates based on a pure dataset, where items exhibiting DIF are removed, are compared with estimates based on the original dataset. Further analyses are necessary to investigate the extent and importance of DIF across the language versions.

There is no gold standard method of assessing MID and so several methods are often used. Different distribution-based statistics are available. However, these different approaches usually produce different magnitudes of MID. In addition, the results often differ to those obtained using anchor-based estimates (Turner et al, 2010). Anchor-based methods are usually given more weight when estimating the MID as they relate scores to other meaningful measures. However, these estimates also have limitations. If a comparator measure is used it is important for it to be adequately related to the PRO being studied (Puhan et al, 2008; Schunemann et al, 2003). In addition, global change items are frequently used but little is known about the reliability of these assessments (Guyatt et al, 2002; Norman et al, 1997).

In estimating the MID other considerations should also be made. Previous research has suggested that MID may be different for patients with different levels of severity (Hajiro et al, 2002; Guyatt et al, 2002). The PRIMUS and U-FIS study investigated the MID in a fairly mild sample of patients with relapsing remitting MS. The MID may need to be reinvestigated in different MS samples. In addition, it is possible that the MID varies depending on whether a patient improves or deteriorates (Cella et al, 2002; Kwok et al, 2010; Colangelo et al, 2009). In the present study there was a bi-directional difference for the U-FIS with individuals who improved having a larger MID than those who deteriorated. This was not found for the PRIMUS scales.

5.5 Chapter summary

The ALIS and PRIMUS were adapted successfully into several new language versions. These new versions are now available for international research studies. Both adaptations used a unique dual panel methodology for the translation. This methodology is widely used and is the only one applied to adaptations of needs-based measures. The PRIMUS adaptation showed how Rasch analysis can be used to improve psychometric evaluation of adapted language versions. Adaptations for both measures would benefit from further analysis to assess for DIF by language.

The MID study was successful in providing estimates for the PRIMUS and U-FIS. The estimates will help interpretation of scores and sample size determination for future trials. The MID values provided are specific to patients with relapsing remitting MS. Further analyses may be necessary to determine MID estimates for other forms of MS.

Chapter 6: Evaluation of existing patient-reported outcomes

6.1 Introduction and articles

Three studies are discussed that describe the evaluation of established PRO measures. The PROs are evaluated in relation to their construct definition and statistical methodology.

6.1.1 Article 1: Validation of the mood disorder questionnaire for screening for bipolar disorder in a UK sample

Twiss J, Jones S, Anderson I. Validation of the Mood Disorder Questionnaire for screening for bipolar disorder in a UK sample. J Affect Disord. 2008 Sep;110(1-2):180-4. doi: 10.1016/j.jad.2007.12.235.

BACKGROUND: The Mood Disorder Questionnaire (MDQ) was designed as a screening questionnaire for bipolar disorder. Previous research has raised questions about the suitability of the MDQ structure for screening for bipolar II disorder. This study investigated the optimal sensitivity and specificity cut-off thresholds for the MDQ in bipolar I and bipolar II patients in a UK sample.

METHODS: The MDQ was administered to patients before attending a tertiary mood disorders clinic. Diagnostic interviews were used to determine DSM-IV diagnoses and these were used as the gold standard against which to investigate the performance of the MDQ.

RESULTS: 54 patients with bipolar spectrum disorder and 73 patients with unipolar depressive disorder completed the MDQ. With the original scoring criteria (symptoms and supplementary questions) the sensitivity for bipolar disorder was 0.76 (bipolar I disorder 0.83, bipolar II disorder 0.67) with specificity 0.86. The optimal cut-off score in the current sample was a score of 9 or more endorsed symptoms without applying the supplementary questions (sensitivity of 0.90 and 0.88 for bipolar I and bipolar II groups respectively with a specificity of 0.90).

LIMITATIONS: The sample was drawn from a tertiary mood disorders clinic.

CONCLUSIONS: The MDQ appears to be a useful screening tool for bipolar spectrum disorder in UK psychiatric practice with sensitivity for bipolar II disorder improved by dropping the supplementary sections. Further investigation of the optimal cut-off scores of the MDQ is needed to determine its utility in non-specialist and community based samples.

The article can be accessed via: http://www.jad-journal.com/article/S0165-0327(08)00016-5/abstract

6.1.2 Article 2: Can we rely on the dermatology life quality index as a measure of the impact of psoriasis or atopic dermatitis?

Twiss J, Meads DM, Preston EP, Crawford SR, McKenna SP. Can we rely on the Dermatology Life Quality Index as a measure of the impact of psoriasis or atopic dermatitis? J Invest Dermatol. 2012 Jan;132(1):76-84. doi: 10.1038/jid.2011.238.

The Dermatology Life Quality Index (DLQI) is a widely used health-related quality of life measure. However, little research has been conducted on its dimensionality. The objectives of the current study were to apply Rasch analysis to DLQI data to determine whether the scale is unidimensional, to assess its measurement properties, test the response format, and determine whether the measure exhibits differential item functioning (DIF) by disease (atopic dermatitis versus psoriasis), gender, or age group. The results show that there were several problems with the scale, including misfitting items, DIF by disease, age, and gender, disordered response thresholds, and inadequate measurement of patients with mild illness. As the DLQI did not benefit from the application of Rasch analysis in its development, it is argued that a new measure of disability related to dermatological disease is required. Such a measure should use a coherent measurement model and ensure that items are relevant to all potential respondents. The current use of the DLQI as a guide to treatment selection is of concern, given its inadequate measurement properties.

The article can be accessed via:
http://www.nature.com/jid/journal/v132/n1/full/jid2011238a.html

6.1.3 Article 3: Psychometric performance of the CAMPHOR and SF-36 in pulmonary hypertension

Twiss et al. BMC Pulmonary Medicine 2013, 13:45
http://www.biomedcentral.com/1471-2466/13/45

BMC Pulmonary Medicine

RESEARCH ARTICLE Open Access

Psychometric performance of the CAMPHOR and SF-36 in pulmonary hypertension

James Twiss[1*], Stephen McKenna[1], Louise Ganderton[2,3,4,5], Sue Jenkins[3,4,6], Mitra Ben-L'amri[1], Kevin Gain[2,4,7], Robin Fowler[2,3,4] and Eli Gabbay[2,3,4,7,8]

Abstract

Background: The Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) and the Medical Outcomes Study Short Form 36 (SF-36) are widely used to assess patient-reported outcome in individuals with pulmonary hypertension (PH). The aim of the study was to compare the psychometric properties of the two measures.

Methods: Participants were recruited from specialist PH centres in Australia and New Zealand. Participants completed the CAMPHOR and SF-36 at two time points two weeks apart. The SF-36 is a generic health status questionnaire consisting of 36 items split into 8 sections. The CAMPHOR is a PH-specific measure consisting of 3 scales; symptoms, activity limitations and needs-based QoL. The questionnaires were assessed for distributional properties (floor and ceiling effects), internal consistency (Cronbach's alpha), test-retest reliability and construct validity (scores by World Health Organisation functional classification).

Results: The sample comprised 65 participants (mean (SD) age = 57.2 (14.5) years; n(%) male = 14 (21.5%)). Most of the patients were in WHO class 2 (27.7%) and 3 (61.5%). High ceiling effects were observed for the SF-36 bodily pain, social functioning and role emotional domains. Test-retest reliability was poor for six of the eight SF-36 domains, indicating high levels of random measurement error. Three of the SF-36 domains did not distinguish between WHO classes. In contrast, all CAMPHOR scales exhibited good distributional properties, test retest reliability and distinguished between WHO functional classes.

Conclusions: The CAMPHOR exhibited superior psychometric properties, compared with the SF-36, in the assessment of PH patient-reported outcome.

Background

Pulmonary hypertension (PH) is associated with progressive elevation of pulmonary artery pressure (PAP) and pulmonary vascular resistance (PVR), leading to right ventricular failure and premature death [1]. Pulmonary arterial hypertension is a rare condition with an estimated incidence of 2-7 per million per year [2,3]. However, incidence rates are considerably higher when other subtypes of PH are considered [4]. Previous research has indicated a higher prevalence in females of around 1.5 to 3 times that of men [3]. PH presents with nonspecific symptoms, including dyspnea on exertion, fatigue and syncope. These symptoms are often difficult to separate from those caused by other disorders, leading to late diagnosis [5]. Patients can experience severe limitations in physical activity requiring lifestyle modifications [6] and the inability to maintain employment [7]. The psychological impact of PH can result in social isolation, depression [8-10] and diminished quality of life [11].

Several types of outcome measure are available for determining the impact of PH. Haemodynamic variables, such as PVR, are often used as primary endpoints in clinical trials. However, evidence shows that these do not correlate well with the impact of the illness from the patients' perspective [12]. Measures of physical function, such as the 6-minute walk distance (6MWD), are also frequently used. Although these measures provide objective data they do not capture the impact of the disease on patients. Researchers often use patient-reported outcome measures (PROMs) to determine the wider impact of PH from the patient's perspective.

There are two main types of PROMs; generic and disease-specific. Generic outcome measures are used with a wide range of illnesses. These measures are popular as

* Correspondence: jtwiss@galen-research.com
[1]Galen Research Ltd, Manchester, United Kingdom
Full list of author information is available at the end of the article

they are thought to negate the need to develop a new measure for each disease studied. One limitation of generic measures is that they may not assess concerns that are unique to each illness and important to patients. Disease-specific measures are developed to assess the specific concerns of the patient group [13].

The two most widely used PROMs with PH patients are the Medical Outcomes Study Short-Form 36 general health survey (SF-36) [14] and the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) [15]. The SF-36 is a generic health-related quality of life (HRQL) measure that has been used in several clinical trials for PH. Despite this, limited information is available regarding the psychometric properties of the SF-36 in a PH population. Previous research has shown that the SF-36 correlates with functional measures such as the 6MWD and New York Heart Association assessment of functional class [12]. In addition, there is some evidence that the SF-36 is responsive in the PH population [16]. However, findings have been inconsistent and only some of the SF-36 domains appear to be responsive [17-19]. In addition, the investigation of scores representing the minimal important difference (MID) of the SF-36 in this patient group has shown that some of the domains of the SF-36 have large MID values [20]. This implies that large changes in scores are required to indicate a real change in health status.

The CAMPHOR is a PH-specific measure and comprises three scales assessing impairments (symptoms), activity limitations (functioning) and quality of life (QoL). A further development of the measure led to a utility scale for use in economic evaluations [21]. The content for the measure was derived directly from patient interviews and embodies issues important to patients with PH. The CAMPHOR has been shown to have good construct validity and reproducibility [15]. All three scales have been shown to fit the Rasch model providing evidence of unidimensionality. In addition, there is evidence that the scales are responsive to change [22]. Although the psychometric properties of the CAMPHOR are promising, direct comparisons with other measures are lacking.

The aim of this study was to conduct a direct comparison of the psychometric properties of the CAMPHOR and the SF-36 in a single population of PH patients in order to determine the suitability of each as an outcome measure.

Methods

Participants

The study utilizes data collected in Australia and New Zealand [23]. Participants were men and women over the age of 18 years, who met World Health Organisation (WHO) [24] criteria for the diagnosis of PH. Participants were required to be native English speaking and were excluded if they were unable to complete the questionnaires due to cognitive impairment. Ethics committees at Royal Perth Hospital and Curtin University in Australia gave approval for the study. Informed consent was obtained from the participants.

Outcome measures

CAMPHOR

The CAMPHOR was developed in the United Kingdom (UK) [15] and subsequently adapted for use in Australia and New Zealand [23]. It consists of three scales; the Symptom Scale and QoL Scale both consist of 25 items with a dichotomous response format (Yes/No). Scores can range from 0-25 with a low score indicating minimal symptoms or better QoL. The Activity Scale consists of 15 items with a 3 point rating system (Able to do on own without difficulty/Able to do on own with difficulty/Unable to do on own). Scores range from 0-30 with a low score indicating minimal activity limitation.

SF-36; version 2

The SF-36 [14] is a generic health status questionnaire consisting of eight domains; physical functioning (10 items), social functioning (2 items), role limitations due to physical problems (4 items), role limitations due to emotional problems (3 items), mental health (5 items), energy/vitality (4 items), pain (2 items), general health perception (5 items) and a single health transition item. Raw domain scores are transformed to a scale of 0-100 with high scores indicating better health status.

Procedure

Details of the methodology are reported in full elsewhere [23]. In brief, the study was conducted via postal survey. Participants completed the SF-36 and CAMPHOR at two time-points, two weeks apart. They also provided demographic and disease information (age, gender, WHO class and PH type). Participants completed the SF-36 immediately followed by the CAMPHOR at each time point (Time 1 [T1] and Time 2 [T2]).

Statistical analyses

Data were analysed using SPSS Version 16.0. Data are provided for T1 and T2 assessment points throughout the results section.

Distributional properties

The distributional properties of the CAMPHOR and SF-36 were examined using descriptive statistics including mean, standard deviation, median, inter-quartile range and range. The proportion of participants scoring the minimum and maximum possible scores on the questionnaires was also assessed. This provides an indication of the targeting of the questionnaire to the patient group. A

high proportion of participants scoring at the extremes can indicate lack of sensitivity and/or relevance.

Internal consistency

Internal consistency was assessed using Cronbach's alpha coefficients for CAMPHOR and SF-36. This coefficient measures the extent to which items in a scale are inter-related. A low alpha (below 0.7) indicates insufficient relations between the items to form a scale [25].

Test-retest reliability

The test-retest reliability of a measure is an estimate of its reproducibility over time when no change in the condition being assessed has taken place. The test-retest reliability of the CAMPHOR and the SF-36 was examined by correlating scores collected at T1 and T2 using Spearman's rank correlation coefficients. A correlation coefficient greater than or equal to 0.85 is required to indicate that a scale has low random measurement error [26]. It is important to note that the Spearman's correlation coefficient does not represent the percentage of explained variance. To assist with the interpretation of the correlation coefficient, the percentage of variance explained in the CAMPHOR and SF-36 scores (r^2) was calculated. In addition, corresponding confidence intervals for mean scores were provided based on the standard error of measurement (SEM) to indicate the level of accuracy inherent in the scores. The SEM is useful for estimating how participants may score during repeated applications of the same measure. Confidence intervals based on the SEM show how participants' scores are distributed around their 'true scores'. Measures with lower reliability will have higher SEM values and wider confidence intervals. The SEM is defined in terms of the standard deviation (δ) and the reliability (r) as follows:

$$SEM = \delta\sqrt{(1-r)}$$

Construct validity (Known group validity)

Construct validity was determined using non-parametric tests for independent samples (Mann-Whitney U Test) to test for differences in CAMPHOR and SF-36 scores between groups according to disease severity (WHO functional classification). A p value of <0.05 was considered statistically significant.

Results

Descriptive statistics

Sixty-five participants (51 females, 78.5%) were recruited to the study. Demographic information for the sample is shown in Table 1.

Table 1 Demographics of the study subjects (n=65)

Gender	
Male (%)	14 (21.5)
Female (%)	51 (78.5)
Age	
Mean (SD)	57.2 (14.5)
Median (IQR)	57.8 (47.5-67.8)
Range	20.1-87.5
WHO Classification	
I (%)	3 (4.6)
II (%)	18 (27.7)
III (%)	40 (61.5)
IV (%)	4 (6.2)
PH Type	
Idiopathic PAH (%)	37 (56.9)
Familial PAH (%)	1 (1.5)
Associated PAH (%)	23 (35.4)
Chronic thromboembolic PH (%)	2 (3.1)
PH associated with lung diseases (%)	2 (3.1)

Distributional properties

Total score descriptive information for the SF-36 is shown in Table 2. Results indicated that there were high levels of ceiling effects (% scoring maximum) for the bodily pain, social functioning and role-emotional domains of the SF-36 at both T1 and T2.

Total scale score descriptive information for the CAMPHOR is shown in Table 3. Minimal levels of floor and ceiling effects were found at each time point indicating the scales were well matched to the disease severity levels of the participants.

Internal consistency

The Cronbach's alpha coefficients for the SF-36 and CAMPHOR are shown in Table 4. Values were acceptable (>0.70) for all scales for both measures. This indicates that items are sufficiently related to form scales.

Test-retest reliability

Test-retest reliability, confidence intervals for mean scores and percentage of explained variance for the SF-36 and CAMPHOR are shown in Table 5. Test-retest reliability was good for the SF-36 physical functioning and general health domains. Test-retest correlations were below 0.85 for all other SF-36 domains. These SF-36 domains also had wide confidence intervals for mean scores (indicating score inaccuracy) and had low levels of explained variance ($r^2 < 0.70$).

Test-retest coefficients were good for all CAMPHOR scales, indicating low levels of random measurement error.

Table 2 Descriptive statistics for SF-36 domains

	Physical functioning	Role-physical	Bodily pain	General health	Vitality	Social functioning	Role-emotional	Mental health
Time 1								
n	**59**	**60**	**61**	**60**	**61**	**61**	**60**	**61**
Median (IQR)	35.0 (20.0-50.0)	37.5 (0.0-67.5)	52.0 (41.0-74.0)	30.0 (15.0-47.0)	37.5 (18.8-59.4)	75.0 (37.5-87.5)	75.0 (50.0-97.5)	65.0 (52.5-85.0)
Mean (SD)	35.3 (21.9)	41.9 (37.9)	53.4 (25.1)	30.3 (19.8)	38.2 (23.8)	62.1 (31.1)	67.9 (41.2)	67.4 (17.9)
Range	0.0-80.0	0.0-100.0	0.0-100.0	0.0-77.0	0.0-81.3	0.0-100.0	0.0-100.0	30.0-100.0
Floor effect (% scoring min)	5.1	5.0	4.9	8.3	6.6	3.3	6.7	3.3
Ceiling effect (% scoring max)	3.4	3.3	9.8	3.3	4.9	21.3	25.0	1.6
Time 2								
n	**59**	**60**	**60**	**58**	**61**	**60**	**61**	**61**
Median (IQR)	30.0 (20.0-50.0)	40.6 (12.5-73.4)	51.5 (33.5-74.0)	25.0 (13.8-42.8)	31.3 (18.8-53.1)	62.5 (37.5-87.5)	75.0 (50.0-95.8)	70.0 (55.0-85.0)
Mean (SD)	35.2 (21.7)	42.6 (28.0)	54.5 (26.5)	30.3 (21.3)	37.1 (21.3)	60.8 (29.4)	68.0 (28.3)	70.0 (16.4)
Range	0.0-90.0	0.0-100.0	0.0-100.0	0.0-87.0	0.0-81.3	0.0-100.0	0.0-100.0	25.0-100.0
Floor effect (% scoring min)	6.8	6.7	3.3	3.4	4.9	3.3	3.3	1.6
Ceiling effect (% scoring max)	1.7	1.7	13.3	1.7	1.6	20.0	24.6	3.3

Table 3 Descriptive statistics for CAMPHOR scales

Time 1	Symptoms	Activities	QoL
n	**65**	**65**	**65**
Median (IQR)	14.0 (7.0 – 18.5)	9.0 (5.0 – 14.5)	11.0 (4.0 – 16.0)
Mean (SD)	13.0 (6.0)	9.9 (5.9)	10.4 (6.5)
Range	2.0 – 23.0	0.0 – 24.0	0.0 – 23.0
Floor effect (% scoring min)	0.0	3.1	6.2
Ceiling effect (% scoring max)	0.0	0.0	0.0
Time 2			
n	**65**	**65**	**65**
Median (IQR)	11.0 (7.0 – 17.0)	10.0 (6.0 – 15.0)	12.0 (5.0 – 16.0)
Mean (SD)	12.5 (6.0)	10.8 (6.1)	10.8 (6.3)
Range	1.0 – 25.0	0.0 – 23.0	0.0 – 23.0
Floor effect (% scoring min)	0.0	4.6	3.1
Ceiling effect (% scoring max)	1.5	0.0	0.0

In addition, the confidence intervals were narrow and the scales had high levels of explained variance (Table 5).

Construct validity - Known group validity

Known group validity results are shown in Table 6 and 7. Several of the SF-36 domains distinguished between participants based on their WHO functional classification. However, the bodily pain and mental health domains did not discriminate between groups at either time point (Table 6). The role-emotional domain discriminated between groups at T1 but not T2 (Table 6).

The CAMPHOR was able to discriminate between participants based on WHO functional classification groups (I&II and III&IV) at T1 and T2. Significantly higher scores were found for WHO groups III and IV (Table 7).

Discussion

This study compared the psychometric properties of two widely used PROMs for patients with PH. The results of the study showed that the CAMPHOR had excellent psychometric properties while weaknesses were apparent in several of the SF-36 domains.

Table 4 Cronbach's alpha coefficients for the SF-36 and CAMPHOR

		Time 1	Time 2
SF-36	Physical functioning	.90	.90
	Role-physical	.94	.95
	Bodily pain	.92	.93
	General health	.74	.77
	Vitality	.88	.86
	Social functioning	.91	.85
	Role-emotional	.94	.91
	Mental health	.78	.81
CAMPHOR	Symptoms	.89	.89
	Activities	.91	.91
	QoL	.91	.91

Participants were predominantly in WHO classes II and III indicating moderately severe disease. Despite this three of the eight SF-36 domains (social functioning, role emotional and bodily pain) had high ceiling effects suggesting the participants in this study had no health problems. It is clear these domains lack sensitivity for this patient group. This could be due to the scales containing too few items (2-3 items each). It is also possible that the content of the items is not relevant to this patient group.

Six of the eight SF-36 domains demonstrated inadequate test-retest reliability ($r<0.85$).Two additional statistics were included to assist with interpreting this finding; the percentage of explained variance and standard error of measurement. The SF-36 domains that did not meet acceptable levels of reliability explained only 49-66% of variance in scores. These domains also had high SEM values and wide confidence intervals. Taken together, this indicates that six of the eight SF-36 domains had high levels of random measurement error and inaccuracy. The low reliability of these SF-36 domains suggests that these are not acceptable as a measure intended for use in clinical trials and other types of research in individuals with PH, where the ability to measure changes over time is important. Only the SF-36 physical functioning and general health domains met the required criteria in this sample. In contrast, all of the CAMPHOR domains met the test-retest criteria and showed low levels of random measurement error. This indicates that, unlike the SF-36 outcome, a change in CAMPHOR score is more likely to represent a real change in clinical condition and/or QoL.

Table 5 Test-retest reliability and explained variance

		Test-retest	% of explained variance (r^2)	Time 1 mean	SEM	Corresponding confidence intervals
SF-36	Physical Functioning	.93	86	35.3	5.8	29.5-41.1
	Role-Physical	.81	66	41.9	12.2	29.7-54.1
	Bodily Pain	.72	52	53.4	13.3	40.1-66.7
	General Health Perceptions	.94	88	30.3	4.9	25.5-35.1
	Vitality	.78	61	38.2	11.2	27.0-49.4
	Social Functioning	.76	58	62.1	15.2	46.9-77.3
	Role-Emotional	.70	49	67.9	17.1	50.8-85.0
	Mental Health	.75	56	67.4	9.0	58.5-76.4
CAMPHOR	Symptoms	.86	74	13.0	2.2	10.8-15.2
	Activities	.87	76	9.9	2.1	7.8-12.0
	QoL	.94	88	10.4	1.6	8.8-12.0

Several of the SF-36 domains were able to distinguish between WHO functional classification groups. However, the bodily pain and mental health domains did not distinguish between groups at either time point and the role-emotional domain did not distinguish between groups at Time 2. Although the social functioning scale distinguished between groups the differences in scores failed to reach the thresholds published for the MIDs for this patient group [20]. These findings raise further doubts about the suitability of these domains of the SF-36 for use with this patient group. Emotional symptoms are important features of PH. It is likely that the role-emotional section is not specific enough to PH to measure the construct adequately.

A recent study by Matura et al [27] in the US associated CAMPHOR and SF-36 scores with symptom clusters in PH patients. They found that severity of symptoms was related to outcomes on both measures. However, they did not explore the psychometric performance of the measures. It was interesting to note that scores on the psycho-social domains of the SF-36 (as in the present study) were remarkably high.

Other researchers have investigated the functioning of the SF-36 physical (PCS) and mental (MCS) component summaries in PH patients [28]. Chen et al reported low levels of end effects for the MCS and PCS scales. Considerable doubt has been raised about the validity of the statistical methodology employed in the calculation of these scales [29-36]. Both the PCS and MCS scores are calculated by using factor coefficients from all eight domains. The PCS includes positively weighted coefficients from the physical domains of the measure but also negatively weighted coefficients from the mental domains. This means that in order to obtain the highest PCS scores it is necessary to both have high scores on the physical domains and low scores on the mental domains. The same is true of the MCS. Such an approach to measurement leads to anomalies, including the creation of artificially low end effects. Therefore it was decided not to report PCS or MCS scores in the present study.

Based on the findings of this study only the SF-36 physical functioning and general health perceptions domains met adequate psychometric criteria for use

Table 6 Mean (SD) SF-36 scores by WHO functional classification

	Physical functioning	Role-physical	Bodily pain	General health	Vitality	Social functioning	Role-emotional	Mental health
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Time 1 (n)	**59**	**60**	**61**	**60**	**61**	**61**	**60**	**61**
WHO Classification								
I and II	49.5 (21.6)	61.3 (26.9)	60.1 (25.9)	38.9 (18.5)	48.2 (20.5)	75.0 (27.4)	82.9 (19.3)	72.1 (17.1)
III and IV	27.9 (18.4)	31.4 (22.5)	49.9 (24.3)	25.7 (19.1)	33.0 (24.0)	55.3 (31.1)	59.8 (33.5)	64.9 (18.0)
***p* value**	<.001	<.001	.360	.009	.015	.014	.010	.165
Time 2 (n)	**59**	**60**	**60**	**58**	**61**	**60**	**61**	**61**
WHO Classification								
I and II	48.5 (23.8)	59.4 (20.7)	61.7 (25.8)	38.3 (20.7)	47.9 (20.9)	72.6 (24.2)	76.2 (23.3)	73.8 (16.7)
III and IV	28.3 (17.2)	34.2 (27.6)	50.6 (26.4)	26.4 (20.8)	31.4 (19.4)	54.5 (30.3)	63.8 (30.0)	68.0 (19.1)
***p* value**	.001	.001	.158	.034	.005	.027	.123	.167

p value, Mann-Whitney U-tests.

Table 7 Mean (SD) CAMPHOR scores by WHO functional classification

		Symptoms	Activities	QoL
Time 1	n	Mean (SD)	Mean (SD)	Mean (SD)
WHO classification				
I and II	21	10.4 (5.3)	7.2 (5.7)	7.3 (6.1)
III and IV	44	14.3 (5.9)	11.2 (5.7)	11.9 (6.2)
p value		0.012	0.011	0.007
Time 2				
WHO classification				
I and II	21	10.1 (5.3)	7.7 (6.2)	7.8 (5.8)
III and IV	44	13.6 (6.0)	12.3 (5.5)	12.2 (6.1)
p value		0.031	0.003	0.006

p value, Mann-Whitney U-tests.

in research in individuals with PH. The general health perceptions section of the SF-36 is concerned with perceptions of health and illness beliefs and the physical functioning scale with functional limitations. These outcomes measure only a limited aspect of patients' experience with PH. The results of this study demonstrate that the CAMPHOR is a more complete tool to assess the impact of PH from the patients' perspective, with good psychometric properties in all scales.

As the CAMPHOR is a disease-specific measure the content is highly relevant to PH patients. The low levels of floor and ceiling effects and high test-retest reliability show the measure is sensitive and has low levels of random measurement error. This in turn suggests the CAMPHOR will be responsive to change. A previous research study has provided evidence of the responsiveness of the CAMPHOR [22].

Limitations of the study are noted. A relatively small sample was available so the results should be interpreted with some caution (n=65). However, this is typical of studies in this orphan disease [16,37,38]. A high proportion of females were included in the sample (78.5%). This reflects the gender ratio prevalence in PH patients [3]. The study was not designed to compare responsiveness of the two measures. Despite this, psychometric analyses suggest that the CAMPHOR scales would be more responsive. Overall, the study has provided a good indication of the psychometric properties of the two measures.

Conclusions

Only the SF-36 physical functioning and general health perceptions domains met adequate psychometric criteria for use in research on individuals with PH. In contrast, all three CAMPHOR scales met the criteria. The CAMPHOR has superior psychometric properties to the SF-36 in the assessment of PH patient-reported outcome.

Competing interests

The present work was unfunded. JT, SPM and MB are employees of Galen Research (GR). GR developed and own the copyright of the CAMPHOR. The other authors have no conflict of interest.

Authors' contributions

JT and SPM were involved in the design of the study. LG, SJ, KG, RF and EG were involved in data acquisition and management. JT, SPM and MB conducted the psychometric evaluation. All authors contributed to the interpretation of the results. The manuscript was drafted by JT and SPM and all authors contributed to its critical review. The final manuscript was approved by all authors.

Acknowledgements

The authors would like to thank the patients for their participation in this study and the following clinicians from Australia and New Zealand for their assistance in recruiting patients: Dr Lutz Beckert (Christchurch Hospital, Christchurch, New Zealand), Dr Fiona Kermeen (The Prince Charles Hospital, Queensland, Australia), Cherie Franks (The Prince Charles Hospital, Queensland, Australia), Dr Eugene Kotlyar (St Vincent's Hospital, New South Wales, Australia), Carolyn Corrigan (St Vincent's Hospital, New South Wales, Australia), Dr Susanna Proudman (Royal Adelaide Hospital, South Australia, Australia), Leah McWilliams (Royal Adelaide Hospital, South Australia, Australia), Professor Trevor Williams (The Alfred, Victoria, Australia) and Cristianne Manterfield (The Alfred, Victoria, Australia).

Author details

[1]Galen Research Ltd, Manchester, United Kingdom. [2]Royal Perth Hospital, Perth, Australia. [3]Lung Institute of Western Australia, Centre for Asthma, Allergy and Respiratory Research, University of Western Australia, Crawley, Australia. [4]School of Physiotherapy and Curtin Health Innovation Research Institute, Curtin University, Perth, Australia. [5]Discipline of Physiotherapy, Faculty of Health Sciences, The University of Sydney, Darlington, Australia. [6]Sir Charles Gairdner Hospital, Perth, Australia. [7]School of Medicine and Pharmacology, University of Western Australia, Perth, Australia. [8]School of Medicine, University of Notre Dame, Fremantle, Australia.

Received: 22 January 2013 Accepted: 3 July 2013
Published: 12 July 2013

References

1. Rubin LJ: **Primary pulmonary hypertension.** *N Engl J Med* 1997, **336**(2):111–117.
2. Rudarakanchana N, Trembath RC, Morrell NW: **New insights into the pathogenesis and treatment of primary pulmonary hypertension.** *Thorax* 2001, **56**(11):888–890.
3. Peacock AJ, Murphy NF, McMurray JJ, Caballero L, Stewart S: **An epidemiological study of pulmonary arterial hypertension.** *Eur Respir J* 2007, **30**:104–109.
4. Strange G, Playford D, Stewart S, Deague JA, Nelson H, Kent A, Gabbay E: **Pulmonary hypertension: prevalence and mortality in the Armadale echocardiography cohort.** *Heart* 2012, **98**(24):1805-11.
5. Rubin LJ: **Diagnosis and management of pulmonary arterial hypertension: ACCP evidence-based clinical practice guidelines.** *Chest* 2004, **126**(1 Suppl):7S–10S.
6. Wryobeck J, Lippo G, McLaughlin V, Riba M, Rubenfire M: **Psychosocial aspects of pulmonary hypertension: a review.** *Psychosomatics* 2007, **48**:467–475.
7. Rubenfire M, Lippo G, Bodinia B, Blasi F, Allegra L, Bossone E: **Evaluating health-related quality of life, work ability, and disability in pulmonary arterial hypertension.** *Chest* 2009, **136**:597–603.
8. Lowe B, Grafe K, Ufer C, *et al*: **Anxiety and depression in patients with pulmonary hypertension.** *Psychosom Med* 2004, **66**:831–836.
9. McCollister DH, Beutz M, McLaughlin V: **Depressive symptoms in pulmonary arterial hypertension: prevalence and association with functional status.** *Psychosomatics* 2010, **51**(4):339–339. e8.

10. Kendler KS, Karkowski LM, Prescott CA: **Causal relationship between stressful life events and the onset of major depression.** *Am J Psychiatry* 1999, **156**(6):837–841.
11. Cenedese E, Speich R, Dorschner L, Ulrich S, Maggiorini M, Jenni R, Fischler M: **Measurement of quality of life in pulmonary hypertension and its significance.** *Eur Resp J* 2006, **28**:808–815.
12. Chua R, Keogh AM, Byth K, O'Loughlin A: **Comparison and validation of three measures of quality of life in patients with pulmonary hypertension.** *Intern Med J* 2006, **36**(11):705–10.
13. McKenna SP: **Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science.** *BMC Med* 2011, **14**(9):86.
14. Ware JE, Kosinski M, Dewey JE: *How to score version two of the SF-36 health survey.* QualityMetric, Incorporated: Lincoln, RI; 2000.
15. McKenna SP, Doughty N, Meads DM, Doward LC, Pepke-Zaba J: **The Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR): a measure of health-related quality of life and quality of life for patients with pulmonary hypertension.** *Qual Life Res* 2006, **15**(1):103–15.
16. Souza R, Jardim C, Martins B, Cortopassi F, Yaksic M, Rabelo R, Bogossian H: **Effect of bosentan treatment on surrogate markers in pulmonary arterial hypertension.** *Curr Med Res Opin* 2005, **21**(6):907–11.
17. Pepke-Zaba J, Gilbert C, Collings L, Brown MC: **Sildenafil improves health-related quality of life in patients with pulmonary arterial hypertension.** *Chest* 2008, **133**(1):183–9.
18. Mok MY, Tsang PL, Lam YM, Lo Y, Wong WS, Lau CS: **Bosentan use in systemic lupus erythematosus patients with pulmonary arterial hypertension.** *Lupus* 2007, **16**(4):279–85.
19. Wong RC, Koh GM, Choong PH, Yip WL: **Oral sildenafil therapy improves health-related quality of life and functional status in pulmonary arterial hypertension.** *Int J Cardiol* 2007, **119**(3):400–2.
20. Gilbert C, Brown MC, Cappelleri JC, Carlsson M, McKenna SP: **Estimating a minimally important difference in pulmonary arterial hypertension following treatment with sildenafil.** *Chest* 2009, **135**(1):137–42.
21. McKenna SP, Ratcliffe J, Meads DM, Brazier JE: **Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses.** *Health Qual Life Outcomes* 2008, **6**:65.
22. Meads DM, McKenna SP, Doughty N, Das C, Gin-Sing W, Langley J, Pepke-Zaba J: **The responsiveness and validity of the CAMPHOR utility index.** *Respir J* 2008, **32**(6):1513–1519.
23. Ganderton L, Jenkins S, McKenna SP, Gain K, Fowler R, Twiss J, Gabbay E: **Validation of the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) in Australian and New Zealand populations.** *Respirology* 2011, **16**:1235–1240.
24. Simonneau G, Galiè N, Rubin LJ, Langleben D, Seeger W, Domenighetti G, Gibbs S, Lebrec D, Speich R, Beghetti M, Rich S, Fishman A: **Clinical classification of pulmonary hypertension.** *J Am Coll Cardiol* 2004, **43**:5S–12S.
25. Streiner D, Norman G: *Health measurement scales.* Oxford: Oxford University Press; 1989.
26. Weiner EA, Stewart BJ: *Assessing individuals.* Boston: Little Brown; 1984.
27. Matura LA, McDonough A, Carroll DL: **Cluster analysis of symptoms in pulmonary arterial hypertension: a pilot study.** *Eur J Cardiovasc Nurs* 2012, **11**(1):51–61.
28. Chen H, De Marco T, Kobashigawa EA, Katz PP, Chang VW, Blanc PD: **Comparison of cardiac and pulmonary specific quality-of-life measures in pulmonary arterial hypertension.** *Eur Respir J* 2011, **38**:608–616.
29. Simon GE, Revicki DA, Grothaus L, Vonkor M: **SF-36 summary scores. Are physical and mental health truly distinct?** *Med Care* 1998, **36**:567–72.
30. Taft C, Karlsson J, Sullivan M: **Do SF-36 summary scores accurately summarise subscale scores?** *Qual Life Res* 2001, **10**:395–404.
31. Wilson D, Parsons J, Tucker G: **The SF-36 summary scales: problems and solutions.** *Soz Praventivmed* 2000, **45**:239–246.
32. Farrivar SS, Cunningham WE, Hays RD: **Correlated physical and mental health summary scores for the SF-36 and SF-12 health survey.** *Health Qual Life Outcomes* 2007, **5**:54.
33. Hann M, Reeves D: **The SF-36 summary scales are not accurately summarized by independent physical and mental component scores.** *Qual Life Res* 2008, **17**:413–23.
34. Agnastopoulos F, Niakis D, Tountas Y: **Comparison between exploratory factor analytic and SEM-based approaches to constructing SF-36 summary scores.** *Qual Life Res* 2009, **18**:53–63.
35. Fleishman JA, Selim AJ, Kasiz LE: **Deriving SF-12 v2 physical and mental health summary scores: a comparison of different scoring algorithms.** *Qual Life Res* 2010, **19**(2):231–41.
36. Tucker G, Adams R, Wilson D: **Observed agreement problems between Sub-scales and summary components of the SF-36 version 2 - an alternative scoring method can correct the problem.** *PLoS One* 2013, **12**:8(4).
37. Strange G, Keogh AM, Williams TJ, Wlodarczyk J, Mcneil KD, Gabbay E: **Bosentan therapy in patients with pulmonary arterial hypertension: the relationship between improvements in 6 minute walk distance and quality of life.** *Respirology* 2008, **13**:674–682.
38. Jing ZC, Yu ZX, Shen JY, Wu BX, Xu KF, Zhu XY, Pan L, Zhang ZL, Liu XQ, Zhang YS, Jiang X, Galiè N, Efficacy and Safety of Vardenafil in the Treatment of Pulmonary Arterial Hypertension (EVALUATION) Study Group: **Vardenafil in pulmonary arterial hypertension: a randomized, double-blind, placebo-controlled study.** *Am J Respir Crit Care Med* 2011, **183**(12):1723–1729.

doi:10.1186/1471-2466-13-45
Cite this article as: Twiss *et al.*: Psychometric performance of the CAMPHOR and SF-36 in pulmonary hypertension. *BMC Pulmonary Medicine* 2013 **13**:45.

6.2 Description of studies

The first article describes a validation study in a UK sample of the Mood Disorder Questionnaire (MDQ; Hirschfeld et al, 2000). The measure was developed to aid the diagnosis of bipolar spectrum disorders and is based closely on the DSM-IV criteria (American Psychiatric Association, 2000). It has three sections; symptom endorsement (Section 1), symptom clustering (Section 2) and severity of problems caused (Section 3). In Section 1 there are thirteen dichotomous items that ask patients whether they have ever experienced different hypomanic symptoms (e.g. "you had much more energy than usual"). Section 2 asks if the symptoms have ever occurred at the same time. Section 3 asks patients about the severity of the symptoms on a four point scale. The original validation study reported that the MDQ performed relatively well against DSM-IV diagnosis (sensitivity 0.73, specificity 0.90) in a psychiatric outpatient population (Hirschfeld et al, 2000). The purpose of this study was to re-assess the functioning of the measure in a UK sample. The results showed that the measure worked well. The optimal cut-off score in the sample was nine or more endorsed symptoms without applying the supplementary questions (sensitivity of 0.90 and 0.88 for bipolar I and bipolar II groups respectively with a specificity of 0.90).

The second article describes the evaluation of the Dermatology Life Quality Index (DLQI) (Finlay and Khan, 1994). This measure is the most frequently used dermatology-specific outcome measure and is used to determine whether patients are eligible to receive biological interventions for psoriasis in the UK (Smith et al, 2005, 2009; NICE, 2008a, 2008b, 2009). It contains ten items covering symptoms, treatment, activity limitations and emotional reactions to having a skin disease. The classical psychometric properties of the questionnaire have been shown to be adequate (Basra et al, 2008). However, more recent analysis using Rasch analysis highlighted several weaknesses with the scale (Nijsten et al, 2006, 2007). The aims of the study were to reassess its measurement properties using Rasch analysis and to determine whether the scale worked in the same way with psoriasis and atopic dermatitis patients. The findings showed that the measure exhibited item misfit, response option dysfunction and differential item functioning (DIF) by disease (psoriasis vs atopic dermatitis). The DIF suggested that some items are interpreted and valued differently in different diseases indicating that scores should not be combined for these two patients groups.

The third study compared the psychometric properties of the SF-36 (Ware et al, 2000) and the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR; McKenna et al, 2006). These two measures are the most frequently used PROs for patients with pulmonary hypertension. The SF-36 is a generic health status questionnaire consisting of eight domains; physical functioning (ten items), social functioning (two items), role limitations due to physical problems (four items), role limitations due to emotional problems (three items), mental health (five items), energy/vitality (four items), pain (two items), general health perception (five items) and a single health transition item. The CAMPHOR is a pulmonary hypertension (PH) specific measure and comprises three scales assessing impairments (symptoms), activity limitations (functioning) and needs-based quality of life (QoL; McKenna and Doward, 2004). The aim of the study was to compare the psychometric properties of the two scales using classical test theory (CTT) analyses. The results showed high ceiling effects (% scoring maximum) for the SF-36 bodily pain, social functioning and role emotional domains indicating a lack of item coverage or lack of suitability of the scales. Test-retest reliability was poor for six of the eight SF-36 domains (Spearman Rank correlation coefficients <0.85), indicating high levels of random measurement error. Three of the SF-36 domains did not distinguish between WHO classes. In contrast, all CAMPHOR scales exhibited good distributional properties, test-retest reliability and distinguished between WHO functional classes.

6.3 Methodology

Ethics approval was gained for the collection of data in each of the three studies. The studies involved secondary analysis of anonymised data.

In the MDQ study the aim was to assess how well the instrument could correctly identify patients with bipolar disorder. The sample included one hundred and twenty seven patients, fifty four with a bipolar spectrum disorder and seventy three with a unipolar diagnosis. Participants were sequential outpatient attendees of a tertiary NHS Specialist Service for Affective Disorders who completed the MDQ and then received a semi-structured clinical interview covering current and past mood disorder (DSM-IV-TR diagnosis; American Psychiatric Association, 2000). The MDQ was evaluated using classical psychometric methods and by ROC curve analysis. Internal reliability was assessed using Cronbach's alpha (Cronbach, 1951). ROC curve analysis was used to assess the measure's sensitivity (proportion of patients with the disease correctly identified) and specificity (proportion of patients without the disease correctly identified). Full diagnostic interview using DSM-IV criteria was used as the gold standard for identifying patient's disease type.

In the DLQI study, one hundred and forty seven patients with psoriasis and one hundred and forty seven patients with atopic dermatitis were included. The evaluation used Rasch analysis to assess the measurement properties of the scale. The analyses conducted were consistent with published guidelines (Tennant and Conaghan, 2007). Several fit indices were assessed including overall fit to the model, individual item fit, response option functioning, coverage of the trait by the items and differential item functioning by disease.

The study comparing the SF-36 and CAMPHOR included data collected in Australia and New Zealand (Ganderton et al, 2011). Sixty five participants completed the SF-36 and CAMPHOR at two time-points, two weeks apart. They also provided demographic and disease information (age, gender, WHO class and PH type). Participants completed the SF-36 immediately followed by the CAMPHOR at each time point. CTT analyses included distributional properties (including % scoring the minimum and maximum), internal consistency (Cronbach's alpha), test-retest reliability and known group validity by WHO functional class.

6.4 Evaluation

The MDQ study represented the first assessment of the functioning of the screening tool in a UK sample. The sensitivity and specificity values were similar to those in the previous research studies in other non-UK samples (Isometsa et al, 2003; Kemp et al, 2008; Miller et al, 2004; Weber Rouget et al, 2005; Benazzi and Akiskal, 2003). The article makes an important contribution to the detection of Bipolar Disorder in the UK.

Given the prominence of the DLQI it is important to have a thorough evaluation of the functioning of the measure. Previous researchers had highlighted some of the deficiencies in the scale (Nijsten et al, 2006, 2007). This study supported these findings and showed the measure worked differently in different disease areas. Importantly, the research suggests that the widespread use of the measure, particularly its use for guiding treatment decisions should be questioned.

The comparison of the SF-36 and CAMPHOR provides a clear illustration of the functioning of each of the measures for patients with pulmonary hypertension. As these are the two most widely used measures in the disease area the study helps researchers identify the most appropriate scale to use.

6.4.1 Construct definition

The MDQ was developed from clinical experts rather than patients. Although no formal conceptual model is specified for the measure, it is based closely on the diagnostic and statistical manual of mental disorders (DSM-IV) criteria for Bipolar Disorder. The DSM-IV lists several different types of Bipolar Disorder including Bipolar I and Bipolar II disorder. Bipolar I disorder is characterised by manic episodes in which patients experience increased arousal and energy levels which may or may not be accompanied by episodes of depression. Bipolar II disorder is characterised by episodes of depression and hypomania, a lesser form of mania. A structured clinical interview for DSM-IV Axis I Disorders (SCID-I) is used for making the diagnoses (First et al, 1996). The criterion requires that patients have experienced a number of manic/hypomanic symptoms together and that the mood issues cause significant distress or impairment of social, occupational or other areas of functioning.

The approach to mental health adopted by the DSM-IV has been criticized for a variety of reasons. It has been argued that there is commonly overlap between different disorders and that distinguishing between conditions such as mania and schizophrenia is challenging (Bentall, 2003). According to this view the strict definitions suggested in the DSM-IV may not reflect the fluidity of mental health issues. Such arguments undermine the validity of some parts of the classification system. In addition, it has been argued that the use of the DSM-IV is leading to the over medicalisation of the general population. For example, the most recent version of the DSM (DSM-V; American Psychiatric Association, 2013) has lowered the threshold for the diagnosis of depression and generalised anxiety disorder. Finally, the content of the DSM-IV was developed via a 'task force' of clinical experts using closed practices which is likely to cause bias (Cosgrove and Regier, 2009). Despite these criticisms the DSM-IV provides a clearly operationalised system for categorising mental health conditions. By adhering closely to the DSM-IV approach the MDQ benefits from the large body of work conducted in this area.

No formal conceptual basis for the DLQI is provided by the authors. The content of the measure was derived from information provided by one hundred and twenty dermatology outpatients. Patients were asked to write down all the ways that their skin condition affected them. The content of the measure was then based on the most frequently mentioned issues. The interpretation of the results and selection of items was not guided by a theoretical underpinning. Failure to describe the conceptual basis of an outcome measure adequately is not acceptable as it means that the validity of the scale must be questioned. It is not surprising that without a clear conceptual basis the DLQI includes a mixture of types of outcomes including symptoms (e.g. Itchy, sore, painful, or stinging), functioning (e.g. Interferes with shopping/looking after home/garden) and other issues such as treatment problems.

The SF-36 has gone through numerous stages of evolution which makes tracking the original conceptual basis challenging. Items used in the SF-36 scales were derived from several different instruments that had been in use for twenty-to-forty years (Ware et al, 1992). A shorter twenty-item SF-36 questionnaire was created first from the larger previous scales (Stewart et al, 1988). It was then decided to lengthen the scale to include thirty six items. The reason for lengthening the measure was an attempt to increase its sensitivity.

The underlying theoretical basis of the measure is not clear and the reason for selecting the particular domains is not explicitly stated. Confusingly different conceptual frameworks for the measure have been suggested in which the domains are combined into different higher order scales such as psychical functioning, general and mental functioning (Keller et al, 1998). As discussed in Chapter 2 (section 2.5.3), conceptual frameworks are limited in their explanation of a construct if they are not embedded in a wider theoretical body of work.

The weaknesses in both these measures have important clinical implications. In clinical practice, it is not clear how clinicians or practitioners should interpret scores on the measures making their use very limited. In addition, as it is not clear what each PRO measures, results may be misinterpreted in clinical trials.

In contrast the CAMPHOR scales were based on clearly defined conceptual foundations. The impairments (symptoms) and activity limitations (functioning) scales are based on the World Health Organization's (WHO) classification of impairments (physiological and anatomical) and activity limitations (capacity and performance) (World Health Organisation, 1980, 1999). The QoL scale utilizes the needs-based approach to quality of life (McKenna and Doward, 2004). Despite this, the construct definition for the three CAMPHOR scales could be improved by defining their underlying measurement mechanisms (Stenner et al, 2013). Further work is necessary to identify clearly how each of the measures work and what governs the location of the items on the underlying construct.

6.4.2 Psychometric methods

ROC curve analysis was used to assess the most appropriate cut-off values for the MDQ. The results showed that the scale had greater levels of sensitivity and specificity if only the first section covering symptom reporting was used. A cut off value of nine provided the greatest level of sensitivity and specificity. This contrasts with the original study which used the second and third qualifying questions and identified a cut off score of seven. The ROC curve analyses were appropriate for the assessment and allowed thorough investigation of the performance of the screening tool.

Cronbach's alpha coefficients were also calculated for the symptoms section and found to be adequate (Alpha coefficient = 0.91; Item–item total correlations ranged from 0.41–0.81). Unfortunately, Cronbach's alpha is limited in its ability to inform on the measurement properties of a scale and the alpha value can be artificially increased by including similar items (Streiner, 2003). Assessment using Rasch analysis would have provided greater detail on the measurement properties of the scale and allowed the assessment of unidimensionality, item fit and coverage of the underlying trait.

Unidimensionality was not assessed in the original development of the DLQI (Finlay and Khan, 1994). Subsequently the measurement properties have been assessed using a two-parameter form of IRT and suggested that fit was adequate (Mazzotti et al, 2006). Unfortunately, this analysis was lacking in detail and firm conclusions cannot be drawn from the data presented. Furthermore the Rasch model is a more powerful IRT model as it provides measurement that has specific objectivity and sufficiency (Stenner, 1994; Linacre, 1992). These are key requirements of fundamental measurement and allow interval level measurement to be achieved.

The DLQI showed overall misfit, item level misfit, response option dysfunction, poor measurement range and DIF by disease. The combination of these results raises concerns about the suitability of the DLQI for guiding treatment decisions as is the case in the UK. The DLQI has been described as a first generation health related quality of life (HRQL) measure (Nijsten et al, 2007) as it was developed without the use of modern psychometric methods and without a clear conceptual foundation.

Classical psychometric analyses were used to compare the functioning of the CAMPHOR and SF-36 due to the sample limitations caused by the rarity of the disease. It would have been desirable to assess the functioning of the measures using Rasch analysis. The CAMPHOR was developed using Rasch analysis and each of its scales were found to fit to the model (McKenna et al, 2006). The SF-36 was not developed using Rasch analysis and the scaling properties of the measure are unclear. Many of the scales are too short for proper assessment of these properties. For example, the bodily pain and social functioning scales contain only two items. The physical functioning scale contains ten items and the measurement properties of the scale have been investigated in several studies. The results of these studies have been mixed. Two papers have indicated misfit to the Rasch model (Haley et al, 1994; McHorney et al, 1997). Other studies have suggested that the physical functioning scale shows better fit (Taylor and McPherson, 2007). Evidence of DIF by disease has also been identified in other studies suggesting that scores may not be comparable in different diseases (Dallmeijer et al, 2007).

6.5 Chapter summary

In this chapter three studies evaluating the psychometric properties of existing scales have been considered. Each study makes an important contribution to knowledge in their respective fields. The screening tool for Bipolar disorder has provided an effective way for the disease to be screened in everyday practice in the UK. The evaluation of the DLQI exposed limitations of the measure. The results suggest that the DLQI may not be suitable for making treatment decisions. Finally, the comparison of the two most widely used PROs in pulmonary hypertension showed the relative strengths of each scale which is essential information for researchers and clinicians selecting the most useful outcome measure.

Each study was evaluated using the criteria discussed in Chapters 2 and 3. Weaknesses in the development methods of the scales were evident. In particular, the SF-36 and DLQI have poorly defined constructs and weak measurement properties. These measures can be considered first generation outcomes and new outcomes adopting modern psychometric standards are necessary in order to improve the validity and quality of measurement in these areas.

Chapter 7: Co-calibrating disease-specific patient reported outcomes

7.1 Introduction and article

In this chapter one article is discussed that shows an innovative method for placing two different disease-specific PROs onto the same scale so that scores can be combined and compared across diseases. The process uses Rasch analysis (Rasch, 1960/1980) to co-calibrate the scales onto the same measurement continuum. Two dermatology-specific outcome measures are co-calibrated in the study; the Psoriasis Quality of Life Scale (PSORIQoL; McKenna et al, 2003) and the Quality of Life in Atopic Dermatitis Scale (QoLIAD; Whalley et al, 2004). The method used will be described and its benefits discussed. The importance of the conceptual basis and psychometric properties of the scales are also discussed.

7.1.1 Comparing the impact of psoriasis and atopic dermatitis on quality of life: co-calibration of the PSORIQoL and QoLIAD

Twiss J, McKenna SP. Comparing the impact of psoriasis and atopic dermatitis on quality of life: co-calibration of the PSORIQoL and QoLIAD. Qual Life Res. 2015 Jan;24(1):105-13. doi: 10.1007/s11136-014-0630-y.

BACKGROUND: Disease-specific patient-reported outcome (PRO) measures are designed to be highly relevant to one disease. It is widely believed that comparisons of outcomes between patients with different diseases are only possible using generic measures. The present study employs a novel method of using Rasch analysis to co-calibrate scores from different disease-specific PRO measures, allowing scores to be compared across diseases.

METHODS: Psoriasis patients (n = 146, mean age = 44.4, males = 50 %) completed the Psoriasis Quality of Life scale (PSORIQoL) and atopic dermatitis patients (n = 146, mean age = 45.5, males = 50 %) the Quality of Life in Atopic Dermatitis scale (QoLIAD). Both measures employ the needs-based model of QoL, and they share five common items-providing a link between assessments. The groups were analysed separately, and then combined to test fit to the Rasch model.

RESULTS: Both scales showed good fit to the Rasch model after minor adjustments (PSORIQoL: χ (2) p = 0.25; QoLIAD: χ (2) p = 0.51). For the combined dataset, one common item showing differential item functioning by disease was removed and fit to the Rasch model was achieved (χ (2) p = 0.08). The co-calibrated scale successfully distinguished between perceived severity groups ($p < 0.001$).

CONCLUSIONS: It is possible to co-calibrate scores on the PSORIQoL and QoLIAD. This is one of the first studies in health research to demonstrate how Rasch analysis can be used to make comparisons across diseases using different disease-specific measures. Such an approach maintains the greater relevance and, consequently, accuracy associated with disease-specific measurement.

The article can be accessed via: http://link.springer.com/article/10.1007%2Fs11136-014-0630-y

7.2 Description

In dermatological research individuals with different skin conditions are often combined (Potocka et al, 2008, 2009; Ludwig et al, 2009; Quandt et al, 2008; Papoutsaki et al, 2007; Schmitt et al, 2007). When this is done generic PROs are often used, as their content is not specific to one condition. However, as discussed in Chapter 6, generic outcomes often lack the sensitivity of disease-specific measures (Twiss, 2013; Shikiar et al, 2006; Angst et al, 2008; Dawson et al, 2012). In addition, older generic outcomes often have poor psychometric properties (Twiss, 2012). An alternative method of assessing outcomes for patients with different diseases was discussed in the article. The method used Rasch analysis to co-calibrate different disease-specific measures onto the same measurement scale.

This method has been used frequently in educational settings to equate tests of different difficulty levels and to standardise tests results from one year to another (Wright, 1993). For example, students of different ability levels may sit different forms of a maths test. In order to provide grades to the students the different forms of the maths tests must be placed onto the same measurement scale. Using this approach it is possible for patients with different diseases to be compared when they have completed different disease-specific patient reported outcomes (PROs).

There are two commonly used approaches to co-calibrating different outcome measures (Vale, 1986). These are:

1. Common person design. In this method participants complete both forms of the test and then the measures are co-calibrated. The tests do not have overlapping item content but must measure the same construct.

2. Common Item design. Here participants complete only one form of a test but the two tests have item content that is common to both measures. Again, it is essential that the different tests measure the same construct.

Previous research has co-calibrated different disease-specific PROs using a common person design (Latimer et al, 2012; Thissen et al, 2011; Crane et al, 2008). A clear limitation of this method is that patients must complete both PROs. This may require patients to complete a disease-specific PRO that is not relevant to their disease. Alternatively a common item design can be used. In order to conduct a common item co-calibration it is necessary for different PROs to have the same conceptual foundation and have overlapping item content. Outcome measures based on the needs-based approach to quality of life (QoL) fulfil these requirements.

The PSORIQoL is a psoriasis-specific measure of QoL and the QoLIAD is an atopic dermatitis-specific measure of QoL. Both diseases affect the skin but differ based on factors such as areas affected, itchiness, age of onset, triggers, and associated disorders (O'Neill et al, 2011; Bowcock and Cookson, 2004). Each measure was developed based on the needs-based model of QoL (McKenna and Doward, 2004). Due to this the PSORIQoL and QoLIAD share a number of common items.

The results of the Rasch analyses showed that it was possible to co-calibrate the two scales. The co-calibrated scale showed good measurement properties across a range of analyses. In addition, the scale was able to distinguish between severity groups providing further evidence of validity.

Samples of the PSORIQoL and QoLIAD are provided in appendices 7 and 8. Copies of the full measures can be obtained from gr@galen-research.com.

7.3 Methodology

Secondary analyses were conducted on available data. Ethics approval was sought and obtained in the original studies. Informed consent was gained from the patients and all data anonymised.

The sample consisted of two hundred and ninety two participants from the UK (one hundred and forty seven with psoriasis and one hundred and forty seven with atopic dermatitis). Psoriasis patients completed the PSORIQoL and atopic dermatitis patients completed the QoLIAD. Both scales were developed using the Rasch model and contain five common items. The responses to the common items were used as anchors for the co-calibration analysis.

Three stages were involved in the co-calibration:

1. The PSORIQoL and QoLIAD were analysed separately to test fit to the Rasch model. This was used to establish whether the measurement properties of the scales were suitable for co-calibration.

2. Data from both PROs were then combined using a common item design. Rasch analysis fit statistics for the common items and overall scale were investigated.

3. The validity of the method was then investigated by relating the co-calibrated scores to disease type and perceived disease severity.

7.4 Evaluation

This is one of the first studies to apply Rasch analysis to co-calibrate two disease-specific PROs using a common item design. The study has clear advantages over approaches that use generic outcomes. As the measures are specific to each disease they are more relevant to patients in each group. This should ensure that the outcomes are more sensitive to change than generic measures (Twiss, 2013; Shikiar et al, 2006; Angst et al, 2008; Dawson et al, 2012). Generic measures frequently used in dermatology have weak measurement properties (Nijsten et al, 2006; 2007; Twiss et al, 2012). In addition, when generic outcomes are used in different diseases differential item functioning (DIF) by disease is also a problem (Dallmeijer et al, 2007; Taylor and McPherson, 2007; Jenkinson et al, 2001). This research shows that items are valued differently in different diseases so that scores may not be comparable.

Practically, co-calibration has a clear application in research studies combining patients with different diseases that currently use generic PROs. In addition, this method may also be useful in comparative effectiveness studies that are used to make decisions about the allocation of scarce health resources (Chalkidou and Anderson, 2009). The relative effectiveness of treatment interventions in different conditions can be compared in these situations using a common QoL metric. However, further research is needed to compare the effectiveness of this method with standard generic PROs and to test the usefulness of the method in comparative effectiveness studies.

As a large number of disease-specific outcomes have been developed based on the needs-based approach there is potential to co-calibrate scales across a number of different disease areas.

7.4.1 Construct definition

In order to co-calibrate disease-specific outcomes using a common item design it is essential that the scales measure the same construct. In the present study the needs-based approach was used. This has been applied in a large number of different diseases and is the most widely applied approach in QoL research (McKenna and Doward, 2004). The approach has been described in more detail in Chapters 2, 4 and 5. Although each disease impacts on patients' ability to meet their needs in different ways there is commonly overlap as needs are universal. For example, satisfaction of social needs may be restricted in a range of conditions including rheumatoid arthritis, multiple sclerosis and Crohn's disease. The strength of this approach is that it can be applied across a wide range of different conditions.

Although several outcomes have been developed based on the needs-based approach few other outcomes have been developed using a common conceptual foundation. The clinical application of the co-calibration method is currently limited for this reason.

7.4.2 Psychometric methods

Rasch analysis was used to co-calibrate the two scales. The process of co-calibration is possible due to the ability of the Rasch model to handle missing data. Ability/item parameters can still be estimated with missing responses present. This means two scales can be calibrated onto the same scale where only a proportion of the items have been completed by both samples.

One of the limiting factors in this approach is that many scales do not fit the Rach model. Substantial manipulations, such as item removal and restructuring the response format, may be necessary in order make the scales fit. In many cases even after these manipulations fit to the model may be unsatisfactory. Ultimately, lack of fit to the Rasch model will preclude the application of this method for many PROs.

In the present study five items were available initially for the co-calibration. One of the items exhibited DIF by disease and had to be deleted leaving four items as anchors. There is no consensus in the literature on how many common items are needed for co-calibration. However, the general view is that the more items available the more robust the item calibrations will be (Vale, 1986; Wolfe, 2000). Research has also shown that successful co-calibration can be achieved with relatively few items if the common and unique items are of good enough quality to ensure good estimates of ability (Wingersky and Lord, 1984). In future needs-based instrument development additional emphasis will be placed on selecting items that overlap with existing scales.

7.5 Chapter summary

This study included in this chapter showed that it was possible to co-calibrate two different disease-specific PROs using Rasch analysis. Both PROs were based on the same conceptual foundation: the needs-based approach to QoL. As both measures have overlapping item content a common item design was used. This method is likely to be most applicable in research studies which combine patients with different diseases where generic outcomes are usually used. In addition, the approach may be suitable for comparative effectiveness studies. A large number of PROs are available based on the needs-based approach to QoL making co-calibration across several different diseases possible.

Limitations in the method were identified. These relate to the need for different PROs to have the same conceptual foundation and the requirement that the scales fit the Rasch model. Few available scales are based on a strong theoretical foundation and also fit the Rasch model. These limitations could be overcome if a wider approach to PRO measurement is adopted where disease-specific PRO measures in different disease areas are developed based on the same conceptual foundation.

Chapter 8 – Summary and conclusions

8.1 Overview

The thesis has presented 10 research studies concerned with the improvement of PRO measurement. The research has covered a wide range of topics:

- The development of new PROs
- Application of PROs in international research
- The evaluation of existing PROS
- Co-calibration of disease-specific PROs

An underlying aim of all the research was to improve the standards of measurement in PRO research. Two common themes run through the work:

- The importance of clear PRO construct definition
- High quality psychometric measurement methods

This chapter examines the contribution of the research, reviews the themes, highlights areas for future research and considers whether the aims of the research have been met.

8.2 Contribution of the research

Chapter 4 described the development of 3 new disease-specific PROs. Each of the measures makes a valuable contribution to outcome measurement within their relevant disease area. All of the measures were based on a clear construct definition and applied Rasch analysis in their development. Consequently, the measures achieve a high quality of measurement. The scales are valuable in clinical trials and for monitoring patients' progress in clinical practice.

Chapter 5 described the adaptation of two measures into a number of additional languages using a unique dual panel approach. One of the studies exhibited a method of cross-cultural validation that included Rasch analysis. The development of the adaptations allows the measures to be used in international clinical trials and research studies increasing their practical value.

Chapter 5 also discussed the estimation of the minimal important difference (MID) for the Patient-Reported Indices for Multiple Sclerosis (PRIMUS; Doward et al, 2009b) and the Unidimensional Fatigue Impact Scale (U-FIS; Meads et al, 2009). This study was successful in providing estimates. These will help interpretation of scores to determine whether or not an intervention is effective and for sample size determination.

In Chapter 6, three widely used PROs were evaluated based on the criteria specified in Chapters 2 and 3. The Dermatology Life Quality Index (DLQI) (Finlay and Khan, 1994) and the SF-36 (Ware et al, 2000) were found to have several limitations. The findings have important implications for the way in which the measures are used. Several scales of the SF-36 were shown to be unsuitable in pulmonary hypertension due to high ceiling effects, poor reliability and poor construct validity. The Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR; McKenna et al, 2006) provided more sensitive measurement with less measurement error. The DLQI also showed weaknesses including misfit to the Rasch model, response option dysfunction and differential item functioning (DIF) by disease. The weak psychometric properties of the DLQI are concerning as the measure is currently used to guide treatment decisions in the UK. The results of this study suggest that using the measure in this way may lead to poor clinical decision making.

In Chapter 7, a new method for co-calibrating scores from two disease-specific PROs was discussed. This method offers a way of combining PRO data from patients with different diseases that complete different disease-specific measures. Ordinarily, generic outcomes would be used for this purpose. As discussed in Chapter 6, generic outcomes often lack the sensitivity of disease-specific measures and the older generic measures suffer from poor measurement properties (Twiss, 2013; Nijsten et al, 2006; 2007; Dallmeijer et al, 2007; Taylor and McPherson, 2007; Jenkinson et al, 2001). This method has the potential to provide more accurate and sensitive data when combining patients with different diseases.

8.3 Themes of the thesis

8.3.1 Clear construct definition

Clear construct definition provides the cornerstone of a measure and gives its rationale. Although it is fundamental to good, purposeful measurement it is often not considered adequately by the developers of PROs (Gimeno-Santos et al, 2011; McKenna, 2011a). This part of the measure development is perhaps the most challenging. It requires a thorough understanding of different types of health outcome constructs and which is required for the specific study. It also demands a detailed justification and explanation for the construct. There is no simple way of assessing whether a measure has captured the intended construct adequately (Cano and Hobart, 2011). Evidence must be sought through several approaches including face, internal, content and construct validity assessments.

In Chapter 2 three different approaches to construct definition were discussed. These included clearly defining the theoretical foundation of the construct, providing a conceptual framework for the outcome measure and producing a measurement mechanism for the construct.

The approach to construct definition in each of the PROs included in the thesis has been evaluated. Evaluation of the DLQI and SF-36 showed limitations in the definition of the constructs measured by each. The DLQI includes different types of outcome and does not have a clear theoretical foundation. The SF-36 was developed based on previous outcome measures rather than being based on a clear underlying theory. A conceptual framework for the SF-36 has been provided that shows the items group into different domains. However, the content of the SF-36 is combined into different kinds of outcome depending on which scoring method is applied. Furthermore, there is little justification for the domains selected due to the lack of theoretical underpinning. Due to these limitations it is not clear exactly what each intends to measure. Unfortunately, this is frequently found in older PROs as the conceptual foundation of the measures was not considered adequately (Gimeno-Santos et al, 2011).

There are serious consequences related to not clearly defining the underlying construct of a PRO. In clinical practice it is vital that each outcome measure used should have a clear clinical purpose. This helps clinicians and practitioners to monitor properly the patient's condition. If a PRO is not developed based on a clear theoretical foundation then it will not help clinicians and practitioners to understand the patient's experience and/or could mislead them into making the wrong decisions for the patients care. In addition, the use of such measures in clinical trials can also provide misleading findings. Due to this, patients may not receive appropriate interventions for their condition.

In contrast, needs-based QoL measures provide a clearer theoretical basis for the construct they assess. They define QoL in relation to the satisfaction of human needs and are supported by a large body of research on human motivation (Maslow 1970; Max-Neef et al 1991; Kenrick et al 2010). The development methodology clearly identified needs affected by each condition and this guided item selection.

Despite this, further explanation of this construct is necessary in order to provide a measurement mechanism. Attempts are needed to explain how specific components can be manipulated to allow items to represent different levels of the construct (Stenner et al, 2013). This will lead to greater quality in measurement.

8.3.2 Psychometric measurement approach

Preference has been given in this thesis for the application of Rasch analysis (Rasch, 1960/1980) for developing and evaluating PROs. The main strength of the Rasch model is its embodiment of fundamental measurement. When data fit the Rasch model they achieve interval level measurement. Other forms of IRT do not provide fundamental measurement so were not considered in the research.

However, Rasch analysis is not without its detractors and the model tends to divide IRT researchers into those for and against. Its detractors consider the model to be overly restrictive and not reflective of data produced by most PROs (Ghaemi, 2011). The model is considered overly restrictive as it provides only one parameter; a difficulty parameter. This means that all items share the same level of discrimination. However, it is this restriction in the model that allows fundamental measurement to be achieved. When an additional discrimination parameter is added as with the two and three parameter IRT models, the chance of interval measurement is lost as measurement invariance cannot be achieved.

The difference between the approaches taken using the Rasch model and those using two and three parameter models are subtle but important. One important difference is due to general IRT approaches attempting to fit a model to the data whereas the Rasch approach fits data to the pre-defined model (Andrich, 2004). In a two-parameter model each item is allowed to have a different level of discrimination so a model that best describes the data is selected. In contrast, the Rasch model is defined a priori so data are tested for adequacy given model requirements. Justification of poor measures using IRT models that do not provide fundamental measurement is an unsatisfying methodological approach (Andrich, 2004).

8.4 Limitations of the thesis

8.4.1 Areas not covered

It has not been possible to discuss all of the important aspects of PRO measurement within the confines of this thesis. One important component of PRO development not discussed in detail is PRO design. This area includes the design of the instructions, time reference for the items, item design and format of the response options. Much research has been conducted into each of these components (Tanur 1992; Stull et al 2009; Schneider et al, 2013; Streiner and Norman, 1989; Khadka et al, 2012). This is an area of great importance as it forms a means by which the conceptual basis of the measure is realised and data is collected for the psychometric analyses. Due to this the design aspects of the new PROs included in this thesis were considered carefully.

8.4.2 Methodological limitations

Any research study is guided by the knowledge and practices of the time. In health outcomes research large changes have occurred over the last two decades. Until recently classical test theory was the dominant force in the area. This has been challenged by the emergence of IRT and Rasch analysis (Belvedere and Morton, 2010). These new methods have brought new knowledge and a gradual improvement in the quality of measurement in the field. This gradual progression and improvement is part of the scientific process and the research presented in this thesis is subject to the same gradual shifts.

Knowledge of the practical application of Rasch analysis is advancing as more research is conducted. The statistical analyses conducted as part of the Rasch analyses in this research have been superseded by slightly different techniques. For example, methods for detecting multidimensionality have changed over time. A method that involves comparing estimates from two subsets of items loading most differently on the first residual components analysis is often now used for this purpose (Smith, 2002). This method was not used in the development of some of the earlier needs-based PROs. In addition, new ways of identifying local dependence in the Rasch model are also now used (see Chapter 4, section 4.4.2). Both of these developments may have influenced item selection in some of the needs-based PROs.

8.5 Future research

New measures are being developed based on the needs-based approach to QoL. This includes new measures for Crohn's disease, intestinal failure, ulcerative colitis and neurofibromatosis. These developments will help improve measurement of QoL within these areas, providing high quality measurement of the issues that most affect patients.

The availability of a large number of needs-based QoL measures also offers further opportunity to co-calibrate across several diseases areas as described in Chapter 7. This approach could help to replace the use of older generic outcome measures that are used for this purpose. It could also be applied in comparative effectiveness studies where the relative effectiveness of different therapeutic interventions need to be compared. Research is required to assess whether the process will have value for this purpose.

Developing a clearer understanding of the underlying mechanisms that drive our constructs is also necessary. This will lead to better construct definition and more accurate, purposeful measurement. Future research will attempt to identify the underlying measurement mechanism of the needs-based QoL construct.

References

Allen, M. J. and Yen, W. M. (2002) *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.

American Psychiatric Association. (2000) *Diagnostic and statistical manual of mental disorders (DSM).* 4th ed*.,* Washington, D.C.: American Psychiatric Pub.

American Psychiatric Association. (2013) *Diagnostic and statistical manual of mental disorders (DSM).* 5th ed., Washington, D.C.: American Psychiatric Pub.

Andrich, D. (1988) *Rasch models for measurement.* Newbury Park: Sage Publications Inc.

Andrich, D. (2004) 'Controversy and the Rasch model: a characteristic of incompatible paradigms?' *Medical Care*, 42(1) pp. I7–16.

Andrich, D., Sheridan, B. and Luo, G. (2009) *Interpreting RUMM2030.* 4th ed., Perth: RUMM Laboratory Pty Ltd.

Angst, F., Verra, M. L., Lehmann, S. and Aeschlimann, A. (2008) 'Responsiveness of five condition-specific and generic outcome assessment instruments for chronic pain.' *BMC Medical Research Methodology*, 8(1) pp. 26.

Baghaei, P. 'Local Dependency and Rasch Measures.' (2008) *Rasch Measurement Transactions*, 21 (3) pp. 1105-1106.

Barchard, K. A. and Hakstian, A. R. (1997) 'The effects of sampling model on inference with coefficient alpha.' *Educational and Psychological Measurement*, 57(6) pp. 893-905.

Basra, M. K., Fenech, R., Gatt, R. M., Salek, M. S. and Finlay, A. Y. (2008) 'The Dermatology Life Quality Index 1994-2007: a comprehensive review of validation data and clinical results.' *British Journal of Dermatology*, 159(5) pp. 997-1035.

Belvedere, S. L. and de Morton, N. A. (2010) 'Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments.' *Journal of Clinical Epidemiology*, 63(12) pp. 1287-1297.

Benazzi, F. and Akiskal, H. S. (2003) 'The dual factor structure of self-rated MDQ hypomania: energized-activity versus irritable-thought racing.' *Journal of Affective Disorders*, 73(1) pp. 59–64.

Bentall, R. (2003) *Madness explained: psychosis and human nature*. London: Penguin.

Bowcock, A. M. and Cookson, W. O. (2004) 'The genetics of psoriasis, psoriatic arthritis and atopic dermatitis.' *Human Molecular Genetics*, 1(13) pp. 43-55.

Cano, S. J. and Hobart, J.C. (2001) 'The problem with health measurement.' *Patient Preference and Adherence,* 5, June, pp. 279-290.

Cella, D., Hahn, E. A. and Dineen, K. (2002) 'Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening.' *Quality of Life Research*, 11(3) pp. 207-221.

Chalkidou, K. and Anderson, G. (2009) 'Comparative Effectiveness Research: International Experiences and Implications for the United States.' [Online] [Accessed on 7th March 2015] http://www.nihcm.org/pdf/CER_International_Experience_09.pdf

Chong, H. Y. (2013) 'A Simple Guide to the Item Response Theory (IRT) and Rasch Modelling.' [Online] [Accessed on 12th February 2015] http://www.creative-wisdom.com/computer/sas/IRT.pdf

Christensen, K. G., Engelhard, G. and Salzberger, T. (2012) 'Ask the experts: Rasch vs. factor analysis.' *Rasch Measurement Transactions*, 26(3) pp. 1373-1378.

Christensen, K. B., Kreiner, S. and Mesbah, M. (2013) *Rasch models in health*. UK: ISTE and John Wiley and Sons, Inc.

Colangelo, K. J., Pope, J. E. and Peschken, C. (2009) 'The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36.' *Journal of Rheumatology*, 36(10) pp. 2231-2237.

Cosgrove, L. and Regier, D. A. (2009) *Toward credible conflict of interest policies in clinical psychiatry.* Psychiatric Times. [Online] [Accessed on 29th February 2015] http://www.psychiatrictimes.com/articles/toward-credible-conflict-interest-policies-clinical-psychiatry

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K. and van Belle, G. (2008) 'Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline.' *Journal of Clinical Epidemiology*, 61(10) pp. 1018–1027.

Cronbach, L. (1951) 'Coefficient alpha and the internal structure of tests.' *Psychomerika*, 16(3) pp. 297-334.

Dallmeijer, A. J., de Groot, V., Roorda, L. D., Schepers, V. P., Lindeman, E., van den Berg, L. H., Beelen, A., Dekker, J. and FuPro Study Group. (2007) 'Cross-diagnostic validity of the SF-36 physical functioning scale in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis: a study using Rasch analysis.' *Journal of Rehabilitation Medicine*, 39(2) pp. 163-169.

Dawson, J., Boller, I., Doll, H., Lavis, G., Sharp, R., Cooke, P. and Jenkinson, C. (2012) 'Responsiveness of the Manchester-Oxford Foot Questionnaire (MOXFQ) compared with AOFAS, SF-36 and EQ- 5D assessments following foot or ankle surgery.' *The Journal of Bone and Joint Surgery*, 94(2) pp. 215–221.

De Champlain, A. F. (2010) 'A primer on classical test theory and item response theory for assessments in medical education.' *Medical Education*, 44(1) pp. 109-17.

Donatti, C., Wild, D. and Hareendran, A. (2008) *The use of conceptual models, conceptual frameworks and endpoint models to support label claims of treatment benefit using patient reported outcomes.* ISPOR. [Online] [Accessed on 25th January, 2015] http://www.ispor.org/news/articles/mar08/ucm.asp

Doward, L. C., McKenna, S. P., Meads, D. M., Twiss, J. and Eckert, B. J. (2009b) 'The Development of Patient Reported Outcome Indices for Multiple Sclerosis (PRIMUS).' *Multiple Sclerosis*, 15(9) pp. 1092-1102.

Doward, L. C., McKenna, S. P., Whalley, D., Tennant, A., Griffiths, B., Emery, P. and Veale, D. J. (2009a) 'The Development of the L-QoL: A quality of life instrument specific to Systemic Lupus Erythematosus.' *Annals of the Rheumatic Diseases*, 68(2) pp. 196-200.

Doward, L. C., Gnanasakthy, A. and Baker, M. G. (2010) 'Patient reported outcomes: looking beyond the label claim.' *Health and Quality of Life Outcomes*, 8(1) pp. 89.

Erickson, P., Willke, R. and Burke, L. (2009) 'A concept taxonomy and an instrument hierarchy: Tools for establishing and evaluating the conceptual framework of a patient-reported outcome (PRO) instrument as applied to product labelling claims.' *Value in Health*, 12(8) pp. 1158–1167.

Finlay, A. Y. and Khan, G. K. (1994) 'Dermatology Life Quality Index (DLQI)-a simple practical measure for routine clinical use.' *Clinical and Experimental Dermatology*, 19(3) pp. 210–216.

First, M. B., Spitzer, R. L., Gibbon, M. and Williams, J. B. W. (1996) *Structured clinical interview for DSM-IV axis I disorders, clinician version (SCID-CV)*. Washington, D.C.: American Psychiatric Press, Inc.

Fisk, J. D., Ritvo, P. G., Ross, L., Haase, D. A., Marrie, T. J. and Schlech, W. F. (1994) 'Measuring the functional impact of fatigue: initial validation of the fatigue impact scale.' *Clinical Infectious Diseases*, 18(s1) pp. 79–83.

Frances, A. (2013) 'The new crisis of confidence in psychiatric diagnosis.' *Annals of Internal Medicine,* 159(2) pp. 221–222.

Freal, J. E., Kraft, G. H. and Coryell, J. K. (1984) 'Symptomatic fatigue in multiple sclerosis.' *Archives of Physical Medicine and Rehabilitation*, 65(3) pp. 135–138.

Ganderton, L., Jenkins, S., McKenna, S. P., Gain, K., Fowler, R., Twiss J. and Gabbay, E. (2011) 'Validation of the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) in Australian and New Zealand populations.' *Respirology*, 16(8) pp. 1235–1240.

Gauch Jr, H. G. (2002) *Scientific Method in Practice*. London: Cambridge University Press.

Ghaemi, H. (2011) 'Is Rasch model without drawback? A reanalysis of Rasch model limitations.' *Modern Journal of Language Teaching Methods*, 1(2) pp. 31–38.

Gimeno-Santos, E., Frei, A., Dobbels, F., Rüdell, K., Puhan, M. A., Garcia-Aymerich, J. and PROactive consortium. (2011) 'Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review.' *Health and Quality of Life Outcomes*, 9(1) pp. 86.

Guttman, L. (1950) 'The basis for scalogram analysis.' *In* Stouffer, S. A. (ed.) *Measurement and Prediction*. 4th ed., New York: Wiley,

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W. and Norman, G. R. (2002) 'Methods to explain the clinical significance of health status measures.' *Mayo Clinic Proceedings*, 77(4) pp. 371-383.

Hagell, P., Hedin, P. J., Meads, D. M., Nyberg, L. and McKenna, S. P. (2010) 'Effects of method of translation of patient-reported health outcome questionnaires: a randomized

study of the translation of the Rheumatoid Arthritis Quality of Life (RAQoL) Instrument for Sweden.' *Value in Health*, 13(4) pp. 424-430.

Hajiro, T. and Nishimaru, K. (2002) 'Minimal clinically significant difference in health status: the thorny path of health status measures?' *European Respiratory Journal*, 19(3) pp. 390-391.

Haley, S. M., McHorney, C. A. and Ware, J. E. (1994) 'Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale.' *Journal of Clinical Epidemiology*, 47(6) pp. 671-84.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991) *Fundamentals of item response theory*. Newbury Park, CA: Sage Press.

Hattie, J. (1985) 'Methodology review: Assessing unidimensionality of tests and items.' *Applied Psychological Measurement*, 9(2) pp. 139-164.

Hewlett, S. A. (2003) 'Patients and clinicians have different perspectives on outcomes in arthritis.' *The Journal of Rheumatology*, 30(4) pp. 877-879.

Hobart, J. C., Lamping, D. L., Fitzpatrick, R., Riazi, A. and Thompson, A. J. (2001) 'The Multiple Sclerosis Impact Scale (MSIS-29).' *Brain*, 124(5) pp. 962–973.

Hochster, H. S. (2008) 'The power of "P": On overpowered clinical trials and "positive" results.' *Gastrointestinal Cancer Research*, 2(2) pp. 108-109.

Holland, P. W. and Wainer, H. (1993) *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.

Hirschfeld, R. M. A., Williams, J. B. W., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck Jr., P. E., Lewis, L., McElroy, S. L., Post, R. M., Rapport, D. J., Russell, J. M., Sachs, G. S. and Zajecka, J. (2000) 'Development and validation of a screening instrument for bipolar: the Mood Disorder Questionnaire.' *American Journal of Psychiatry*, 157(11) pp. 1873–1875.

Hunt, S. M. and McKenna, S. P. (1992) 'The QLDS: a scale for the measurement of quality of life in depression.' *Health Policy*, 22(3) pp. 307-319.

IBM Corp. (2011) *IBM SPSS Statistics for Windows, Version 20.0*. Armonk, New York: IBM Corp.

Isometsa, E., Suominen, K., Mantere, O., Valtonen, H., Leppamaki, S., Pippingskold, M. and Arvilommi, P. (2003) 'The mood disorder questionnaire improves recognition of bipolar disorder in psychiatric care.' *BMC Psychiatry,* 3(1) p. 8.

Jenkinson, C., Fitzpatrick, R., Garratt, A., Peto, V. and Stewart- Brown, S. (2001) '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 & Psychiatry*, 71(2) pp. 220–224.

Jones, P. W. (2001) 'Health status measurement in chronic obstructive pulmonary disease.' *Thorax*, 56(11) pp. 880-887.

Keller, S. D., Ware, J. E., Bentler, P. M., Aaronson, N. K., Alonso, J., Apolone, G., Bjorner, J. B., Brazier, J., Bullinger, M., Kaasa, S., Leplège, A., Sullivan, M. and Gandek, B. (1998) 'Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA project.' *Journal of Clinical Epidemiology*, 51(11) pp. 1179-1188.

Kelvin, W.T. (1883) 'Electrical Units of Measurement.' *Popular Lectures*, 1 p. 73.

Kemp, D. E., Hirschfeld, R. M., Ganocy, S. J., Elhaj, O., Slembarski, R., Bilali, S., Conroy, C., Pontau, J., Findling, R. L. and Calabrese, J. R. (2008) 'Screening for bipolar disorder in a county jail at the time of criminal arrest.' *Journal of Psychiatric Research*, 42(9) pp. 778–786.

Kenrick, D. T., Griskevicius, V., Neuberg, S. L. and Schaller, M. (2010) 'Renovating the pyramid of needs: contemporary extensions built upon ancient foundations.' *Perspectives on Psychological Science*, 5(3) pp. 292-314.

Khadka, J., Gothwal, V. K., McAlinden, C., Lamoureux, E. L. and Pesudovs, K. (2012) 'The importance of rating scales in measuring patient-reported outcomes.' *Health and Quality of Life Outcomes*, 10(1) p. 80.

Kline, T. J. B. (2005) *Psychological testing: a practical approach to design and evaluation.* California: SAGE Publications, Inc.

Koch, G. G. (1982) 'Intraclass correlation coefficient.' *In* Kotz, S. and Johnson, N. L. (ed.) *Encyclopedia of Statistical Sciences 4*. New York: John Wiley & Sons, pp. 213–217.

Krupp, L. B., Alvarez, L. A., LaRocca, N. G. and Scheinberg, L. C. (1988) 'Fatigue in multiple sclerosis.' *Archives of Neurology*, 45(4) pp. 435–437.

Krupp, L. B., LaRocca, N. G., Muir-Nash, J. and Steinberg, A. D. (1989) 'The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus.' *Archives of Neurology*, 46(10) pp. 1121–1123.

Kwok, T. and Pope, J. E. (2010) 'Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales.' *Journal of Rheumatology*, 37(5) pp. 1024-1028.

Latimer, S., Covic, T. and Tennant, A. (2012) 'Co-calibration of deliberate self-harm (DSH) behaviours: towards a common measurement metric.' *Psychiatry Research*, 200(1) pp. 26–34.

Lexell, J. E. and Downham, D. Y. (2005) 'How to assess the reliability of measurements in rehabilitation.' *American Journal of Physical Medicine and Rehabilitation,* 84(9) pp. 719–723.

Linacre, J. (1992) 'Why fuss about statistical sufficiency?' *Rasch Measurement Transactions*, 6(3) p. 230.

Linacre, J. M. (1994) 'Sample size and item calibration stability.' *Rasch Measurement Transactions*, 7(4) p. 328.

Ludwig, M. W., Oliveira, M. D. S., Muller, M. C. and Moraes, J. F. (2009) 'Quality of life and site of the lesion in dermatological patients.' *Anais Brasileiros de Dermatologia*, 84(2) pp. 143–150.

Luquet, C., Chau, N., Guillemin, F., Nadif, M., Moreau, F., Gavilott, C. and Pétry, C. (2001) 'A method for shortening instruments using the Rasch model. Validation on a hand functional measure.' *Revue d'Epidémiologie et de Santé Publique*, 49(3) pp. 273–286.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., Snyder, C., Boers, M. and Cella, D. (2011) 'Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting.' *Quality of Life Research*, 21(5) pp. 739-746.

Martin, R. L., Mohtadi, N. G., Safran, M. R., Leunig, M., Martin, H. D., McCarthy, J., Guanche, C. A., Kelly, B. T., Byrd, J. W., Clohisy, J. C., Philippon, M. J. and Sekiya, J. K. (2009) 'Differences in physician and patient ratings of items used to assess hip disorders.' *The American Journal of Sports Medicine*, 37(8) pp. 1508-1512.

Maslow, A. H. (1970) *Motivation and personality.* 2nd ed., New York: Harper & Row.

Max-Neef, M. A., Elizalde, A. and Hopenhayn, M. (1991) *Human scale development: conception, application and further reflections*. New York: The Apex Press.

Mazzotti, E., Barbaranelli, C., Picardi, A., Abeni, D. and Pasquini, P. (2006) 'Reply to the letter by T. Nijsten et al.' *Acta Dermato-Venereologica*, 86(3) pp. 290-291.

McHorney, C. A., Haley, S. M. and Ware, J. E. (1997) 'Evaluation of The MOS SF-36 Physical functioning scale (PF-40): II. Comparison of relative precision using likert and Rasch scoring methods.' *Journal of Clinical Epidemiology*, 50(4) pp. 451-461.

McKenna, S. P., Cook, S. A., Whalley, D., Doward, L. C., Richards, H. L., Griffiths, C. E. and Van Assche, D. (2003) 'Development of the PSORIQoL, a psoriasis-specific measure of quality of life designed for use in clinical practice and trials.' *British Journal of Dermatology*, 149(2) pp. 323-331.

McKenna, S. P. and Doward, L. C. (2004) 'The needs-based approach to quality of life assessment.' *Value in Health*, 7(s1) pp. 1–3.

McKenna, S. P., Doughty, N., Meads, D. M., Doward, L. C. and Pepke-Zaba, J. (2006) 'The Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR): a measure of health-related quality of life and quality of life for patients with pulmonary hypertension.' *Quality of Life Research,* 15(1) pp. 103–115.

McKenna, S. P., Doward, L. C., Twiss, J., Hagell, P., Oprandi, N. C., Fisk, J., Grand'Maison, F., Bhan, V., Arbizu, T., Brassat, D., Kohlmann, T., Meads, D. M. and Eckert, B. J. (2010) 'International Development of the Patient-Reported Outcome Indices for Multiple Sclerosis (PRIMUS).' *Value in Health*, 13(8) pp. 946-51.

McKenna, S. P., Meads, D. M., Doward, L. C., Twiss, J., Pokrzywinski, R., Revicki, D., Hunter, C. J. and Glendenning, G. A. (2011) 'Development and validation of the Living with Chronic Obstructive Pulmonary Disease (LCOPD) Questionnaire.' *Quality of Life Research*, 20(7) pp. 1043-1052.

McKenna, S. P. (2011) 'Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science.' *BMC Medicine*, 9(1) p. 86.

Meads, D. M., Doward, L. C., McKenna, S. P., Fisk, J., Twiss, J. and Eckert, B. (2009) 'The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS).' *Multiple Sclerosis*, 15(10) pp. 1228-1238.

Meads, D. M., McKenna, S. P., Doward, L. C., Pokrzywinski, R., Revicki, D., Hunter, C. and Glendenning, G. A. (2010) 'Development and validation of the Asthma Life Impact Scale (ALIS).' *Respiratory Medicine,* 104(5) pp. 633-643.

Miller, C. J., Klugman, J., Berv, D. A., Rosenquist, K. J. and Ghaemi, N. (2004) 'Sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder.' *Journal of Affective Disorders*, 81(2) pp. 167–171.

Multiple Sclerosis Society UK. (1997) *Symptom management survey multiple sclerosis*. UK: Multiple Sclerosis Society.

Multiple Sclerosis Council for Clinical Practice Guidelines. (1998) *Fatigue and multiple sclerosis: evidence based management strategies for fatigue in multiple sclerosis*. Washington, D.C.: Paralyzed Veterans of America.

National Institute for Health and Clinical Excellence. (2008a) *Infliximab for the treatment of adults with psoriasis*. UK: NICE Technology Appraisal Guidance 134.

National Institute for Health and Clinical Excellence. (2008b) *Adalimumab for the treatment of adults with psoriasis*. UK: NICE Technology Appraisal Guidance 146.

National Institute for Health and Clinical Excellence. (2009) *Ustekinumab for the treatment of adults with moderate to severe psoriasis*. UK: NICE Technology Appraisal Guidance 180.

Neville, C., Clarke, A .E., Joseph, L., Belisle, P., Ferland, D. and Fortin, P. R. (2000) 'Learning from discordance in patient and physician global assessments of systemic lupus erythematosus disease activity.' *The Journal of Rheumatology*, 27(3) pp. 675-679.

Nijsten, T., Meads, D. M. and McKenna, S. P. (2006) 'Dimensionality of the dermatology life quality index (DLQI): a commentary.' *Acta Dermato-Venereologica*, 86(3) pp. 289–290.

Nijsten, T., Meads, D. M., de Korte, J., Sampogna, F., Gelfand, J. M., Ongenae, K., Evers, A. W. and Augustin, M. (2007) 'Cross-cultural inequivalence of dermatology-specific health-related quality of life instruments in psoriasis patients.' *Journal of Investigative Dermatology*, 127(10) pp. 2315-2322.

Nixon, A., Kerr, C., Breheny, K. and Wild, D. (2013) 'Patient Reported Outcome (PRO) assessment in epilepsy: a review of epilepsy-specific PROs according to the Food and Drug Administration (FDA) regulatory requirements.' *Health and Quality of Life Outcomes*, 11(11) p. 38.

Norman, G. R., Stratford, P. and Regehr, G. (1997) 'Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach.' *Journal of Clinical Epidemiology*, 50(8) pp. 869-879.

Novick, M. R. (1996) 'The axioms and principal results of classical test theory.' *Journal of Mathematical Psychology*, 3(1) pp. 1-18.

Nunnally, J. and Bernstein, L. (1994) *Psychometric theory*. New York: McGraw-Hill Higher, Inc.

O'Neill, J. L., Chan, Y. H., Rapp, S. R. and Yosipovitch, G. (2011) 'Differences in itch characteristics between psoriasis and atopic dermatitis patients: results of a web-based questionnaire.' *Acta Dermato-Venereologica*, 91(5) pp. 537-540.

Papoutsaki, M., Chimenti, M. S., Costanzo, A., Talamonti, M., Zangrilli, A., Giunta, A., Bianchi, L. and Chimenti, S. (2007) 'Adalimumab for severe psoriasis and psoriatic arthritis: an open-label study in 30 patients previously treated with other biologics.' *Journal of the American Academy of Dermatology*, 57(2) pp. 269–275.

Pearson, K. (1895) 'Notes on regression and inheritance in the case of two parents.' *Proceedings of the Royal Society of London*, 58, June, pp. 240–242.

Petrillo, J., Cano, S. J., McLeod, L. D. and Coon, C. D. (2015) 'Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples.' *Value in Health*, 18(1) pp. 25-34.

Pett, M. A., Lackey, N. R. and Sullivan, J. J. (2003) *Making sense of factor analysis: the use of factor analysis for instrument development in health care research*. UK: Sage Publications, Inc.

Penner, I. K., Bechtel, N., Raselli, C., Stöcklin, M., Opwis, K., Kappos, L. and Calabrese, P. (2007) 'Fatigue in multiple sclerosis: relation to depression, physical impairment, personality and action control.' *Multiple Sclerosis*, 13(9) pp. 1161–1167.

Popper, K. (1959) *The logic of scientific discovery.* UK: Hutchinson and co.

Popper, K. (1963) *Conjectures and refutations.* UK: Routledge and Kegan Paul.

Potocka, A., Turczyn-Jabłońska, K. and Kieć-Swierczyńska, M. (2008) 'Self-image and quality of life of dermatology patients.' *International Journal of Occupational Medicine and Environmental Health*, 21(4) pp. 309–317.

Potocka, A., Turczyn-Jabłońska, K. and Merecz, D. (2009) 'Psychological correlates of quality of life in dermatology patients: the role of mental health and self-acceptance.' *Acta Dermatovenerologica Alpina, Pannonica et Adriatica*, 18(2) pp. 53–58.

Piquette, C. A., Clarkson, L., Okamoto, K., Kim, J. S. and Rubin, B. K. (2000) 'Respiratory-related quality of life: relation to pulmonary function, functional exercise capacity, and sputum biophysical properties.' *Journal of Aerosol Medicine*, 13(3) pp. 263-72.

Prieto, L., Alonso, J. and Lamarca, R. (2003) 'Classical test theory versus Rasch analysis for quality of life questionnaire reduction.' *Health and Quality of Life Outcomes*, 1(1) p. 27.

Puhan, M. A., Frey, M., Büchi, S. and Schünemann, H. J. (2008) 'The minimal important differences of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease.' *Health and Quality of Life Outcomes*, 6(1) p. 46.

Quandt, S. A., Schulz, M. R., Vallejos, Q. M., Feldman, S. R., Verma, A., Fleischer, A. B., Rapp, S. R. and Arcury, T. A. (2008) 'The association of dermatologist-diagnosed and self-reported skin diseases with skin-related quality of life in Latino migrant farmworkers.' *International Journal of Dermatology*, 47(3) pp. 236–241.

Rasch, G. (1960/1980) Probabilistic models for some intelligence and attainment tests.(Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Raykov, T. (1997) 'Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components.' *Multivariate Behavioural Research*, 32(4) pp. 329 – 353.

Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., Schwartz, C., Revicki, D. A., Moinpour, C. M., McLeod, L. D., Lyons, J. C., Lenderking, W. R., Hinds, P. S., Hays, R. D., Greenhalgh, J., Gershon, R., Feeny, D., Fayers, P.M., Cella, D., Brundage, M., Ahmed, S., Aaronson, N. K. and Butt, Z. (2013) 'ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research.' *Quality of Life Research*, 22(8) pp. 1889-1905.

Revicki, D. A., Osoba, D., Fairclough, D., Barofsky, I., Berzon, R., Leidy, N. K. and Rothman, M. (2000) 'Recommendations on health related quality of life research to support labeling and promotional claims in the United States.' *Quality of Life Research*, 9(8) pp. 887–900.

Revicki, D. A., Gnanasakthy, A. and Weinfurt, K. (2007) 'Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: the PRO Evidence Dossier.' *Quality of Life Research*, 16(4) pp. 717–723.

Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L. and Petrie, C. D. (2009) 'Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report.' *Value in Health*, 12(8) pp. 1075–1083.

Rouget, B. W, Gervasoni, N., Dubuis, V., Gex-Fabry, M., Bondolfi, G. and Aubry, J. M. (2005) 'Screening for bipolar disorders using a French version of the Mood Disorder Questionnaire (MDQ).' *Journal of Affective Disorders*, 88(1) pp. 103–108.

Schmitt, J., Heese, E., Wozel, G. and Meurer, M. (2007) 'Effectiveness of inpatient treatment on quality of life and clinical disease severity in atopic dermatitis and psoriasis vulgaris: a prospective study.' *Dermatology*, 214(1) pp. 68–76.

Schneider, S., Broderick, J. E., Junghaenel, D. U., Schwartz, J. E. and Stone, A. A. (2013) 'Temporal trends in symptom experience predict the accuracy of recall PROs.' *Journal of Psychosomatic Research,* 75(2) pp. 160-166.

Schwid, S. R., Covington, M., Segal, B. M. and Goodman, A. D. (2002) 'Fatigue in multiple sclerosis: current understanding and future directions.' *Journal of Rehabilitation Research and Development*, 39(2) pp. 211–224.

Schunemann, H. J., Griffith, L., Jaeschke, R., Goldstein, R., Stubbing, D. And Guyatt, G. H. (2003) 'Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction.' *Journal of Clinical Epidemiology*, 56(12) pp. 1170-1176.

Scoggins, J. F. and Patrick, D. L. (2009) 'The use of patient-reported outcomes instruments in registered clinical trials: evidence from ClinicalTrials.gov.' *Contemporary Clinical Trials*, 30(4) pp. 289-92.

Shikiar, R., Willian, M. K., Okun, M. M., Thompson, C. S. and Revicki, D. A. (2006) 'The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study.' *Health and Quality of Life Outcomes*, 27(4) p. 71.

Smets, E. M., Garssen, B., Bonke, B. and De Haes, J. C. (1995) 'The Multi-dimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue.' *Journal of Psychosomatic Research*, 39(3) pp. 315–325.

Smith, C. H., Anstey, A. V., Barker, J. N., Burden, A. D., Chalmers, R. J., Chandler, D., Finlay, A. Y., Griffiths, C. E., Jackson, K., McHugh, N. J., McKenna, K.E., Reynolds, N. J. and Ormerod, A. D. (2005) 'British Association of Dermatologists guidelines for use of biological interventions in psoriasis 2005.' *British Journal of Dermatology*, 153(3) pp. 486–497.

Smith, C. H., Anstey, A. V., Barker, J. N., Burden, A. D., Chalmers, R. J., Chandler, D., Finlay, A. Y., Griffiths, C. E., Jackson, K., McHugh, N. J., McKenna, K. E., Reynolds, N. J. and Ormerod, A. D. (2009) 'British Association of Dermatologists guidelines for biologic interventions for psoriasis 2009.' *British Journal of Dermatology*, 161(5) pp. 987–1019.

Smith, E. V. (2002) 'Understanding Rasch measurement: detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.' *Journal of Applied Measurement*, 3(2) pp. 205–231.

Snyder, C. F., Aaronson, N. K., Choucair, A. K., Elliott, T. E., Greenhalgh, J., Halyard, M. Y., Hess, R., Miller, D. M., Reeve, B. B. and Santana, M. (2011) 'Implementing patient-reported outcomes assessment in clinical practice: A review of the options and considerations.' *Quality of Life Research*, 21(8) pp. 1305-1314.

Spearman, C. (1904) 'The proof and measurement of association between two things.' *American Journal of Psychology*, 15(1) pp. 72–101.

Stenner, A. J. (1994) 'Specific objectivity - local and general.' *Rasch Measurement Transactions*, 8(3) p.374.

Stenner, A. J., Burdick, D. S. and Stone, M. H. (2008) 'Formative and reflective models: can a Rasch analysis tell the difference?' *Rasch Measurement Transactions*, 22(1) pp. 1152-1153.

Stenner, A. J., Fisher, W. P., Stone, M. H. and Burdick, D. S. (2013) 'Causal Rasch models.' *Frontiers in Psychology*, 23(4) p. 536.

Streiner, D. L. (2003) 'Starting at the beginning: an introduction to coefficient alpha and internal consistency.' *Journal of Personality Assessment*, 80(1) pp. 99-103.

Streiner, D. L. and Norman, G. R. (1989) *Health measurement scales. A practical guide to their development and use*. UK: Oxford Medical Publications.

Stewart, A. L., Hays, R. D., Ware, J. E. (1988) 'The MOS Short-Form General Health Survey: reliability and validity in a patient population.' *Medical Care*. 26(7) pp. 724-735.

Stull, D. E., Leidy, N. K., Parasuraman, B. and Chassany, O. (2009) 'Optimal recall periods for patient-reported outcomes: challenges and potential solutions.' *Current Medical Research and Opinion*, 24(4) pp. 929–942.

Swaine-Verdier, A., Doward, L. C., Hagell, P., Thorsen, H. and McKenna, S. P. (2004) 'Adapting Quality of Life Instruments.' *Value in health*, 7 (s1) pp. 27-30.

Tanur, J. (1992) *Questions about questions. Inquiries into the cognitive bases of surveys*. New York: Russell Sage Foundation.

Taylor, B. N. (1991) 'The International System of Units (SI): approved translation of the sixth edition (1991) of the International Bureau of Weights and Measures publication Le Système International d'Unités (SI)'. Gaithersburg, MD: National Institute of Standards and Technology.

Taylor, W. J. and McPherson, K. M. (2007) 'Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis.' *Arthritis and Rheumatology*, 57(5) pp. 723-9.

Tennant, A., McKenna, S. P. and Hagell, P. (2004) 'Application of Rasch analysis in the development and application of quality of life instruments.' *Value in Health*, 7(s1) pp. 22–26.

Tennant, A. and Pallant, J. F. (2007) 'DIF matters: a practical approach to test if Differential Item Functioning makes a difference.' *Rasch Measurement Transactions*, 20(4) pp. 1082-1084.

Tennant, A. and Conaghan, P. G. (2007) 'The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one should one look for in a Rasch paper?' *Arthritis and Rheumatism*, 57(8) pp. 1358–1362.

Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E. and DeWalt, D. A. (2011) 'Using the PedsQLTM 3.0 Asthma Module to obtain scores comparable with those of the PROMIS Pediatric Asthma Impact Scale (PAIS).' *Quality of Life Research*, 20(9) pp. 1497–1505.

Traub, R. (1997) 'Classical Test Theory in Historical Perspective.' *Educational Measurement: Issues and Practice*, 16(4) pp. 8-14.

Trojan, D. A., Arnold, D., Collet, J. P., Shapiro, S., Bar-Or, A., Robinson, A., Le Cruguel, J. P., Ducruet, T., Narayanan, S., Arcelin, K., Wong, A. N., Tartaglia, M. C., Lapierre, Y., Caramanos, Z. and Da Costa, D. (2007) 'Fatigue in multiple sclerosis: association with disease-related, behavioural and psychosocial factors.' *Multiple Sclerosis*, 13(8) pp. 985–995.

Turner, R. R., Quittner, A. L., Parasuraman, B. M., Kallich, J. D. and Cleeland, C. S. (2007) 'Patient-reported outcomes: instrument development and selection issues.' *Value in Health*, 10(s2) pp. 86–93.

Turner, D., Schünemann, H. J., Griffith, L. E., Beaton, D. E., Griffith, A. M., Critch, J. N. and Guyatt, G. H. (2010) 'The minimal detectable change cannot reliably replace the minimal important difference.' *Journal of Clinical Epidemiology*, 63(1) pp. 28-36.

Twiss, J., Jones, S. H. and Anderson, I. A. (2008) 'Validation of the Mood Disorder Questionnaire for screening for bipolar disorder in a UK sample.' *Journal of Affective Disorders*, 110(1-2) pp. 180-184.

Twiss, J., Doward, L. C., McKenna, S. P. and Eckert, B. (2010) 'Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS).' *Health and Quality of Life Outcomes*, 8(1) p. 117.

Twiss, J., McKenna, S. P., Crawford, S. R., Oprandi, N. C. and Tammaru, M. (2011) 'Adapting the Asthma Life Impact Scale (ALIS) for use in Southern European (Italian) and Eastern European (Russian) cultures.' *Journal of Medical Economics*, 14(6) pp. 729–738.

Twiss, J., Meads, D. M., Preston, E. P., Crawford, S. R., McKenna, S. P. (2012) 'Can we rely on the Dermatology Life Quality Index (DLQI) as a measure of the impact of psoriasis or atopic dermatitis?' *Journal of Investigative Dermatology*, 132(1) pp. 76-84.

Twiss, J., McKenna, S. P., Ben-L'amri, M., Ganderton, L. and Jenkins, S. (2013) 'Psychometric performance of the CAMPHOR and SF-36 in pulmonary hypertension' *BMC Pulmonary Medicine*, 13 p. 45.

Twiss, J. and McKenna, S. P. (2015) 'Co-calibrating disease specific Quality of life measures for Psoriasis (Psoriasis quality of life measure) and Atopic Dermatitis (Quality of life in Atopic Dermatitis measure).' *Quality of Life Research*, 4(1) pp. 105-13.

US Food and Drug Administration. (2009) 'Patient-reported outcome measures: Use in medical product development to support labeling claims.' Guidance for industry. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/

Vale, C. D. (1986) 'Linking item parameters onto a common scale.' *Applied Psychological Measurement*, 10(4) pp. 333-344.

Vickrey, B. G., Hays, R. D., Harooni, R., Myers, L. W. and Ellison, G. W. (1995) 'A health-related quality of life measure for multiple sclerosis.' *Quality of Life Research*, 4(3) pp. 187–206.

Walters, S. J. and Brazier, J. E. (2005) 'Comparison of the minimally important difference for two health state utility measures: EQ-5 D and SF-6D.' *Quality of Life Research*, 14(6) pp. 1523-1532.

Ware, J. E. and Sherbourne, C. D. (1992) 'The MOS 36-Item Short-FormHealth Survey (SF-36): I. Conceptual framework and item selection.' *Medical Care*, 30(6) pp. 473–483.

Ware, J. E., Kosinski, M. and Dewey, J. E. (2000) *How to score version two of the SF-36 health survey*. Lincoln, RI: QualityMetric, Incorporated.

Wehmeier, P. M., Kluge, M., Schacht, A., Helsberg, K. and Schreiber, W. (2007) 'Correlation of physician and patient rated quality of life during antipsychotic treatment in outpatients with schizophrenia.' *Schizophrenia Research*, 91(1-3) pp. 178-186.

Whalley, D., McKenna, S. P., Dewar, A. L., Erdman, R. A., Kohlmann, T., Niero, M., Cook, S. A., Crickx, B., Herdman, M. J., Frech, F. and Van Assche, D. (2004) 'A new instrument for assessing Quality of Life in Atopic Dermatitis (QoLIAD).' *British Journal of Dermatology*, 150(2) pp. 274–283.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A. and Erikson, P. (2005) 'Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation.' *Value in Health*, 8(2) pp. 94–104.

Wingersky, M. S. and Lord F. M. (1984) 'An investigation of methods for reducing sampling error in certain IRT procedures.' *Applied Psychological Measurement*, 8(3) pp. 347-364.

Wolfe, E. W. (2000) 'Equating and item banking with the Rasch model.' *Journal of Applied Measurement*, 1(4) pp. 409-434.

World Health Organisation. (1980) *The international classification of impairments, disabilities and handicaps*. Geneva: WHO.

World Health Organisation. (1999) *International classification of functioning and disability: ICIDH-2*. Geneva: Stationery Office Books.

Wright, B. D. (1993) 'Equitable test equating.' *Rasch Measurement Transactions*, 7(2) pp.298-299.

Wright, B. D. (1996) 'Comparing Rasch measurement and factor analysis.' *Structural Equation Modelling: A Multidisciplinary Journal*, 3(1) pp. 3–24.

Wright, B. D. and Tennant, A. (1996) 'Sample size again,' *Rasch Measurement Transactions*, 9(4) p. 468.

Wright, B. D. (1997) 'Fundamental Measurement.' *Rasch Measurement Transactions*, 11(2) p. 558.

Waugh, R. F. and Chapman, E. S. (2005) 'An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: what is the difference? Which method is better?' *Journal of Applied Measurement*, 6(1) pp.80–99.

Xitao, F. (1998) 'Item response theory and classical test theory: an empirical comparison of their item/person statistics.' *Educational and Psychological Measurement*, 58(3) p. 357.

Appendices

Appendix 1: Personal contributions to the articles included in the thesis.

Publication	Areas of contribution	% Contribution
Doward, L.C., McKenna, S.P., Meads, D.M., Twiss, J., Eckert, B.J. (2009b). 'The Development of Patient Reported Outcome Indices for Multiple Sclerosis (PRIMUS)'. Multiple Sclerosis, 15(9) pp. 1092-1102.	Qualitative interviews, content analysis, Rasch analysis, classical psychometric analysis, interpretation of results, contribution to drafting of article.	20%
Meads, D., Doward, L., McKenna, S., Fisk, J., Twiss, J., Eckert, B. (2009). 'The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS)'. Multiple Sclerosis, 15 pp. 1228-1238.	Qualitative interviews, content analysis, Rasch analysis, classical psychometric analysis, interpretation of results, contribution to drafting of article.	15%
McKenna, S.P., Meads, D.M., Doward, L.C., Twiss, J., Pokrzywinski, R., Revicki, D., Hunter, C.J., Glendenning, G.A. (2011). 'Development and validation of the Living with Chronic Obstructive Pulmonary Disease (LCOPD) Questionnaire'. Quality of Life Research, 20(7) pp. 1043-1052.	Qualitative interviews, content analysis, classical psychometric analysis, interpretation of results, contribution to drafting of article.	15%
McKenna, S.P., Doward, L.C., Twiss, J., Hagell, P., Oprandi, N.C., Fisk, J., Grand'Maison, F., Bhan ,V., Arbizu, T., Brassat, D., Kohlmann, T., Meads, D.M., Eckert, B.J. (2010). International Development of the Patient-Reported Outcome Indices for Multiple Sclerosis (PRIMUS). Value in Health, 13(8) pp. 946-951.	Rasch analysis, classical psychometric analysis, interpretation of results, contribution to drafting of article.	15%
Twiss, J., McKenna, S.P., Crawford, S.R., Oprandi, N.C., Tammaru, M. (2011). Adapting the Asthma Life Impact Scale (ALIS) for use in Southern European (Italian) and Eastern European (Russian) cultures. Journal of Medical Economics, 14(6), 729–738.	Management, design, analysis, interpretation of results, drafting of article.	75%
Twiss, J., Doward, L.C., McKenna, S.P., Eckert, B. (2010). Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS). Health and Quality of Life Outcomes, 8:117.	Design, analysis, interpretation of results, drafting of article.	80%
Twiss, J., Jones, S.H., Anderson, I.A. (2008). 'Validation of the Mood Disorder Questionnaire for screening for bipolar disorder in a UK sample'. Journal of Affective Disorders, 110(1-2) pp. 180-184.	Design, data analysis, interpretation of results, drafting of article.	70%
Twiss, J., Meads, D.M., Preston, E.P., Crawford, S.R., McKenna, S.P. (2012). 'Can we rely on the Dermatology Life Quality Index (DLQI) as a measure of the impact of psoriasis or atopic dermatitis?'. Journal of Investigative Dermatology, 132(1) pp. 76-84.	Design, data analysis, interpretation of results, drafting of article.	70%

Twiss, J., McKenna, S.P., Ben-L'amri, M., Ganderton, L., Jenkins, S. (2013). 'Comparison of the psychometric properties of the SF-36 and Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for Pulmonary Hypertension patients'. BMC Pulmonary Medicine, 13:45.	Design, data analysis, interpretation of results, drafting of article.	70%
Twiss, J., McKenna, S.P. (2015). Co-calibrating disease specific Quality of life measures for Psoriasis (Psoriasis quality of life measure) and Atopic Dermatitis (Quality of life in Atopic Dermatitis measure). Quality of Life Research, 4(1) pp. 105-13.	Design, data analysis, interpretation of results, drafting of article.	80%

PRIMUS

Patient Reported Indices of Multiple Sclerosis

Please read this carefully

This booklet asks about your experience
of having MS.

Please follow carefully the instructions for each section
and choose the response that best applies to you.

Symptoms

Please read each question carefully and decide whether it has applied to you ***during the last week.*** Put a tick in the box ☑ next to **'Yes'** if you feel it applied to you and a tick in the box ☑ next to **'No'** if it did not.

1. Has your skin been very sensitive? Yes ❒ No ❒

2. Have you experienced weakness in your arms or legs? Yes ❒ No ❒

3. Has your eyesight been blurred? Yes ❒ No ❒

4. Have you had dizzy spells? Yes ❒ No ❒

5. Have you had any muscle spasms? Yes ❒ No ❒

6. Have you had any loss of vision? Yes ❒ No ❒

7. Have you been forgetting things? Yes ❒ No ❒

8. Have you had any numbness? Yes ❒ No ❒

9. Have you had urinary incontinence? Yes ❒ No ❒

10. Have you had bowel incontinence? Yes ❒ No ❒

Appendix 3: Sample of the U-FIS questionnaire

Fatigue Impact Scale (U-FIS)

Below is a list of items that describe the impact of fatigue on people's lives. Please circle the response that best applies to you for each item.

Due to your fatigue, over the last ***week*** how much of the time have you…?

		Never	A little of the time	About half the time	A lot of the time	All the time
1	Run out of energy quickly	0	1	2	3	4
2	Lacked motivation to engage in social activities	0	1	2	3	4
3	Had difficulty dealing with anything new	0	1	2	3	4
4	Found it difficult to organise your thoughts while doing things at home or at work	0	1	2	3	4
5	Found normal day-to-day events stressful	0	1	2	3	4
6	Had to keep stopping and resting	0	1	2	3	4
7	Had difficulty finishing tasks that require thinking	0	1	2	3	4

		Never	A little of the time	About half the time	A lot of the time	All the time
8	Felt you had no energy left for enjoyment/fun	0	1	2	3	4
9	Not felt alert	0	1	2	3	4
10	Had to force yourself to do things	0	1	2	3	4
11	Found it difficult to make decisions	0	1	2	3	4
12	Found that minor difficulties seem like major difficulties	0	1	2	3	4
13	Had difficulty paying attention for a long period of time	0	1	2	3	4
14	Felt unable to meet the demands that people place on you	0	1	2	3	4

Appendix 4: Sample of the LCOPD questionnaire

LCOPD

Quality of life questionnaire

Please read this carefully

On the following pages you will find some statements that
have been made by people who have
Chronic Obstructive Pulmonary Disease (COPD)/breathing problems.

Thinking about your COPD/breathing problems, please read each statement carefully and tick 'True' if the statement applies to you and tick 'Not True' if it does not.

Please choose the response that best applies to you

at the moment.

Remember to tick ☑ the box next to the response that best applies to you at the moment

1. My illness limits the places I can go — True ❒ Not True ❒

2. I get frustrated easily — True ❒ Not True ❒

3. I can't do things on the spur of the moment — True ❒ Not True ❒

4. I feel like a prisoner in my own home — True ❒ Not True ❒

5. I worry that I stop people doing what they want to do — True ❒ Not True ❒

6. My illness controls me — True ❒ Not True ❒

7. I have to plan even the most simple tasks carefully — True ❒ Not True ❒

8. My breathing makes me self conscious — True ❒ Not True ❒

9. I have to pace myself — True ❒ Not True ❒

Appendix 5: Sample of the ALIS questionnaire

ALIS

Quality of life questionnaire

Please read this carefully

On the following pages you will find some statements that have been made by people who have asthma.

Thinking about your asthma, please read each statement carefully and tick 'True' if the statement applies to you and tick 'Not True' if it does not.

Please choose the response that best applies to you

at the moment.

Remember to tick ☑ the box next to the response that best applies to you ***at the moment***

1. Asthma stops me being adventurous	True	❒
	Not True	❒
2. I feel dependent on my treatment	True	❒
	Not True	❒
3. I'm unable to join in activities with my friends and family	True	❒
	Not True	❒
4. I feel older than my years	True	❒
	Not True	❒
5. I have to pace myself	True	❒
	Not True	❒
6. My self-confidence is affected	True	❒
	Not True	❒
7. I constantly have to think about my medication	True	❒
	Not True	❒
8. I have to limit what I do each day	True	❒
	Not True	❒
9. I feel like I let other people down	True	❒
	Not True	❒

Appendix 6: Sample of the CAMPHOR questionnaire

CAMPHOR

Cambridge Pulmonary Hypertension Outcome Review

Please read this carefully

On the following pages you will find some statements that have been made by people who have Pulmonary Arterial Hypertension.

Please read each statement carefully.
We would like you to put a tick in the box ☑ next to **'Yes'** if you feel it applies to you and a tick in the box ☑ next to **'No'** if it does not

Please choose the response that applies best to you

at the moment

Symptoms

Please read each statement carefully and decide whether it applies to you ***at the moment***

1. My stamina levels are low — Yes ❒ No ❒
2. I have to rest during the day — Yes ❒ No ❒
3. I feel worn out — Yes ❒ No ❒
4. I get tired very quickly — Yes ❒ No ❒
5. I'm tired all the time — Yes ❒ No ❒
6. I feel very weak — Yes ❒ No ❒
7. I feel completely exhausted — Yes ❒ No ❒
8. I want to sit down all the time — Yes ❒ No ❒
9. I soon run out of energy — Yes ❒ No ❒

Appendix 7: Sample of the PSORIQoL questionnaire

PSORIQoL

PLEASE READ THIS CAREFULLY

On the following pages you will find some statements
that have been made by people with psoriasis.

We would like you to tick **'True'** if the statement applies to you
and tick **'Not True'** if it does not.

Please choose the response that applies best to you

AT THE MOMENT

Please read each statement carefully and decide whether it applies to you ***at the moment***

1. I worry about what other people think of me
   True ❒ Not True ❒

2. I never feel clean
   True ❒ Not True ❒

3. I hate people seeing my skin
   True ❒ Not True ❒

4. I have no self-confidence
   True ❒ Not True ❒

5. I can't enjoy myself when I go out
   True ❒ Not True ❒

6. Psoriasis rules my life
   True ❒ Not True ❒

Appendix 8: Sample of the QoLIAD questionnaire

QoLIAD

PLEASE READ THIS CAREFULLY

On the following pages you will find some statements
that have been made by people with eczema.

We would like you to tick **'True'** if the statement applies to you
and tick **'Not True'** if it does not.

Please choose the response that applies best to you

AT THE MOMENT

Please read each statement carefully and decide whether it applies to you ***at the moment***

1. I worry about my appearance — True ❒ Not True ❒

2. I have no self-confidence — True ❒ Not True ❒

3. I avoid physical contact — True ❒ Not True ❒

4. I get embarrassed when I am with people I don't know very well — True ❒ Not True ❒

5. My life revolves around my condition — True ❒ Not True ❒

6. I feel tense all the time — True ❒ Not True ❒

Glossary

ANOVA	Analysis of variance
Classical Test Theory (CTT)	An approach to the design, analysis and scoring of tests. It is based on true score theory. The approach uses predominantly correlational based methods and produces measures at the ordinal level of measurement.
Co-calibration	A method of placing two different PROs onto the same measurement scale using Item Response Theory.
Conceptual framework	This explains the structure of a PRO and shows the relation between items, domains and the overall construct measured. It is usually organised in the form of a figure.
Construct	An idea or concept used to explain something. PROs aim to measure different kinds of constructs.
Construct validity	A PRO is considered to have construct validity if it measures what it intends to measure. It is assessed using different approaches such as known group validity.
Convergent validity	This assesses the validity of a PRO by relating it other available outcome measures that assess similar constructs. Higher correlations should be observed between constructs that are more similar.
Content analysis	This involves conducting thematic analysis on interview data. The topic of interest is coded into themes of related issues. Themes are then harmonized until an understanding of the area is developed.
Common person design	A method of co-calibrating scales that requires patients to have completed both of the scales that are to be combined.
Common item design	A method of co-calibrating scales that requires overlap in item content.
Cronbach's alpha coefficient	The Cronbach's alpha coefficient is used to assess the extent to which the items in a scale are inter-related. It is the primary method of assessing internal reliability under Classical Test Theory.
Differential item functioning	This is a statistical method applied in Item Response Theory. It is used to assess whether answers to items are biased by different subgroups (such as those defined by age or gender). This kind of bias causes instability in the severity ordering of the items.
Dual panel translation	A method of translating a questionnaire that uses two translation panels. The first panel consists of group of bilingual speakers that work together to translate the questionnaire. The second panel consists of monolingual speakers of the local language. The role of the second group is to make sure the language selected is easily understood by the target population.
Effect size	This is used to quantify the strength of an observation. It is calculated by dividing the difference between two mean scores by the standard deviation at baseline.

Factor analysis	This includes a group of statistical methods that are used to identify the relations between a set of variables or questionnaire items in order to group them into a smaller number of explanatory domains. The methods are based on correlational techniques.
Forward-backward translation	A method of PRO translation conducted by a linguistic expert. After translation the content of the PRO is translated back into the original source language by a second person. The original and back translated questionnaires are then compared and any discrepancies resolved.
Guttman scale	A Guttman scale is a measure in which the items are ranked in order of difficulty from least extreme to most extreme. Correct answers to the Guttman scale would follow the ordering of the items precisely. A person that answers question 8 correctly would also answer questions 1-7 correctly.
Health related quality of life (HRQL)	An approach to PRO measurement based on assessing different sub-domains relating to a person's health. Such approaches predominantly measure symptoms and functional limitations.
Interval level measurement	Numerical scales where the distances between each part of the scale are the same throughout.
Intra-class correlation	A correlation statistic that accounts for both within-subject change and systematic change in the mean.
Item and person interaction statistics	Used to assess fit to the Rasch model. These assessments measure the extent to which observed item and person estimates deviate from the expected.
Item response theory	Includes a group of models that are concerned with the design, analysis and scoring of tests. Each item is assumed to represents a different level of difficulty. IRT models the response of patients of a given ability to an item of a given difficulty.
Known group validity	This assesses the validity of a PRO by relating it to groups of known importance. For example, scores on the PRO can be related to groups representing different levels of disease severity.
Latent variable	A variable that is not directly observable but is inferred.
Local dependency	A requirement of the Rasch model is the local independence of items. Local dependency occurs when items are too closely related such that the response to one item has too strong an influence over answers to another item.
Measurement mechanism	This is an approach to construct definition whereby the underlying mechanism of the measure is understood so that items can be manipulated to represent varying levels of the construct of interest.
Minimal important difference (MID)	A change score on a measure that represents a minimal level of meaningfulness to the patient.
Needs-based QoL	A definition of quality of life based on the satisfaction of human needs. Quality of life is high when more needs are met.

Ordinal level measurement	A type of measurement where individuals can be ranked but the distances between levels on the scale are unequal.
Overall item-trait interaction Chi^2 fit value	Used to assess overall fit to the Rasch model expectations. A significant Ch^2 value indicates misfit to model expectations.
Patient reported outcome (PRO)	A measure in the form of a questionnaire used to capture information relating to a person's health.
Pearson correlation	A parametric correlation statistic.
Person separation index (PSI)	This is a form of reliability statistic that can be calculated within the Rasch framework and is indicative of the power of the items to distinguish between respondents.
Qualitative interviews	These are open ended interviews in which patients experience with a given topic is explored. The interviews are usually transcribed and then analysed thematically.
Rasch analysis	The Rasch model is a simple logistic one parameter item response theory model with strong mathematical properties. Measures that fit the model provide interval level measurement.
Responder definition	A change score on a measure that represents a minimal level of meaningfulness to the patient. It is also referred to as minimal important difference (MID).
Response threshold	A response threshold is the point between two adjacent response categories where the probability of endorsing either category reaches 0.5. Response thresholds are used within a Rasch framework to assess whether the response options function logically.
Receiving operating characteristic (ROC) curve analysis	A ROC curve analysis is a graphic plot used to assess how well a screening tool classifies individuals at different cut-off levels.
RUMM program	A statistical package used to assess fit to the Rasch model.
Sensitivity analysis	This is used to assess the functioning of a screening tool. It measures the proportion of positives that are correctly identified as such for a given score on a measure.
Specificity analysis	This is used to assess the functioning of a screening tool. It measures the proportion of negatives correctly measured as such for a given score on a measure.
Specification equation	An explanation of the underlying mechanisms that make items represent different levels of the construct of interest.
Spearman Rank correlation	A non-parametric correlation statistic.
Standard error of measurement	This is considered to be an assessment of how much the persons observed score is affected by the error inherent in the test. It is calculated using the standard deviation at baseline and the internal consistency of the measure.

Test-retest reliability	This is a measure of the reproducibility of a questionnaire. A high correlation should be observed between scores on a test when no change in condition has taken place.
True score theory	This is the underlying paradigm of Classical Test Theory. It is based on the assumption that scores on a test are obscured by the error that is inherent in the test. Scores are comprised of 'true score + error'.
Unidimensionality	The property of measuring a single underlying dimension. Measures that fit the Rasch model hold this property.
WHO functional class	A classification of disease severity for pulmonary hypertension. It is clinician completed and comprises four different severity groups.