THE LOWER LIMB TASKS QUESTIONNAIRE: IS IT USEFUL IN THE ASSESSMENT OF FUNCTION IN LOW BACK PAIN SUFFERERS?

ROMA FORBES

A thesis submitted to Auckland University of Technology in partial fulfilment of the requirements for the degree of Master of Health Science (MHSc)

2010

Faculty of Health and Environmental Sciences
Supervisor: Professor Peter McNair and Peter Larmer

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ATTESTATION OF AUTHORSHIP

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

_____

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to the following people:

ABSTRACT

Objective: The objective of this study was to investigate selected measurement properties of the Lower Limb Tasks Questionnaire (LLTQ) in a sample of acute Low Back Pain (LBP) subjects, and compare these findings to the Roland Morris Questionnaire (RMQ). The primary aim of the study was to investigate the responsiveness and minimal clinical important difference (MCID) within this population. Additionally, an analysis of the content and construct validity of the LLTQ was undertaken.

Study Design: A quantitative prospective evaluation of outcome measures, assessing limitations in physical function change over a course of treatment.

Background: Low Back Pain (LBP) is a common condition associated with a loss of function. As restoration of function is a major aim of physiotherapy management and the utilisation of outcome measurements is a fundamental requirement of practice, clinicians must have measures that can detect meaningful change in this construct. The LLTQ is an outcome measure which has been used for populations with lower limb conditions, and in this population it has been shown to demonstrate sound psychometric properties. It offers potential advantages to clinicians and researchers relating to its scoring system, its ability to delineate functional tasks relating to activities of daily life and recreation and to assessing the importance of tasks to the individual.

Methods: Sixty nine subjects who presented for physiotherapy treatment with acute LBP completed the LLTQ and RMQ at the initial visit, when a level of 'improved' had been reached using a 7 point Global Perceived Effect score (GPE), and at discharge. Statistical procedures included, analysis of importance rating data and correlation analysis between baseline LLTQ and RMQ scores to investigate content and construct validity respectively. Responsiveness was estimated using distribution based analyses of effect size, the standardised response mean (SRM) and the minimal detectable change (MDC) for both measures. An anchor based receiver operating characteristic (ROC) curve was generated to establish best cut off points to estimate the MCID of the LLTQ, and was repeated for variations of baseline scores.

Results: The LLTQ was found to cover a wider spread of important functional tasks specifically relating to social, work and employment activities than the RMQ when contrasted with the International Classification of Functioning (ICF) model. A moderate correlation of RMQ scores with both domains of the LLTQ was found (r= .56 and .67), with a significantly stronger correlation demonstrated with the Recreational domain. There was a significant increase in LLTQ scores between the baseline and 'improved' level (p<0.05) and baseline and discharge (p<0.05). The measure demonstrated high levels of responsiveness, with an effect size of 1.6 and 1.7 for the ADL and recreational domains respectively and an SRM of 1.5 for both domains. The MDC was 2.5 and 2.1 points respectively for the ADL and recreational domains. The MCID was 3 points for both domains of the measure, with a likelihood ratio over 10. Further analysis demonstrated a significant relationship between lower baseline scores and higher change scores, however an MCID of 3 points was generated for both domains of the measure regardless of baseline score category.

Conclusion: The results of this study set important benchmarks regarding the ability of the LLTQ to detect both statistical and clinically meaningful change in an acute LBP population. Through contrasting the measure with an external framework of function and the widely employed RMQ, the LLTQ has been shown to have sound content that reflects the limitations of function and priorities of this population. As a practical, valid and responsive measure that can be applied across various clinical populations, the LLTQ thus has the potential to address issues related to the utilisation of outcome measures by clinicians.

CHAPTER 1: INTRODUCTION

1.1 Statement of the Problem

Low back pain (LBP) is currently a major health problem in western societies. In New Zealand, the Accident Compensation Corporation (ACC) has estimated that treatment claims for LBP cost approximately $280 million a year (ACC, 2007). In addition to high economic costs, LBP is associated with loss of function and reduced quality of life for individuals, affecting up to 80% of the population at some point in their lives (Andersson, 1999; Ehrlich, 2003).

Physiotherapy plays an important role in acute LBP management, offering assessment, education and manual and exercise therapy within a primary care setting (Koes, van Tulder & Thomas, 2006; Hayden, van Tulder, Malmivaara & Koes, 2005; Liddle, Baxter & Gracey, 2009). Restoration of physical function is considered one of the key aims of treatment for LBP and is deemed to be one of the most important outcomes for patients, clinicians and funding providers (Beattie & Maher, 1997; Deyo et al, 1998; Grimmer et al, 1999). Although improving physical function is an imperative goal of treatment and patient focussed function outcomes are of primary importance, pain and impairment based measures remain the predominant choice of outcomes used by clinicians despite their questionable relevance, poor validity, reliability and prognostic capacity (Nattrass et al, 1999; Bombardier, 2000; Sullivan, Shoaf & Riddle, 2000).

As part of evidence based practice and professional competency requirements, physiotherapists must employ appropriate outcome measures with particular emphasis on the selection of measures which are patient oriented and have sound measurement properties. Furthermore, clinicians are expected to interpret the outcomes of such measures in order to evaluate the effectiveness of interventions and guide clinical decision making (Australian Council of Physiotherapy Regulating Authorities (ACORPA, 2006); Basmajian, 1995; Fritz & Irrgang, 2001; New Zealand Physiotherapy Board 1999; Resnik & Dobrzykowski, 2003).

In order to assess physical function of an individual and to monitor change over time, clinicians require outcome measures that not only accurately assess function as a construct, but are able to detect change in a valid and reliable way when meaningful

change has occurred, referred to as the responsiveness of a measure. Without knowledge of the level of change in the outcome measure that is deemed meaningful to the patient, the clinician is left to decipher the relevance of the score change. Ascertaining this magnitude of important change is known as the minimal clinically important difference (MCID), and is widely considered to be one of the most important properties of outcome measures (Beaton, Bombardier, Katz & Wright, 2001, Childs, Piva & Fritz, 2005).

To measure limitation in physical function, generic patient reported outcome measures that can be applied to a collective of clinical populations are widely available for use both in clinical and research settings (Deyo et al, 1998; Ostelo et al, 2008; Patrick & Deyo, 1989). There are however, important constraints associated with using such measures (Bombardier, 2000). Generic outcome measures often incorporate physical impairments, quality of life and social and emotional functioning within the same measure, therefore may be insensitive to small, but clinically important change in the construct of physical functioning (Deyo et al, 1998; Resnik & Dobrzykowski, 2003). Such measures are also likely to include functional tasks that relate to various body sites and clinical conditions, resulting in an estimation of function that is neither specific nor sensitive to the true limitation in physical function experienced by the population of interest (Garratt, Klaber Moffett & Farrin, 2001). To address these issues in LBP settings, there have been a considerable number of outcome measures focussed upon function which have been validated and published within recent years (Costa, Maher & Latimer, 2007; Grotle, Brox & Vollestad, 2004; Stier-Jarmer, Cieza, Borchers & Stucki, 2009) yet despite these efforts, the use of outcome measures within physiotherapy practice is not widespread (Horner & Larmer, 2006). The sheer volume of measures available has been suggested as a contributing factor to poor utilisation by clinicians, with other barriers identified including time and financial restraints, lack of knowledge and in some cases, attitudes that such outcome measures are irrelevant (Copeland, Taylor & Dean, 2008). Reports also suggest that therapists have particular concerns regarding administration, scoring, interpretation and documentation (Copeland, Taylor & Dean, 2008; Haigh et al, 2001; Kay, Myers & Huijbrets, 2001; Monk, 2006; Torenbeek, Caulfield, Garrett & van Harten 2001).

The Lower Limb Tasks Questionnaire (LLTQ) possesses several key properties that offer a number of benefits for clinicians and researchers, particularly in assessing

function in those with LBP. The LLTQ consists of two separate domains to assess the individual's limitation in physical function in both activities of daily living (ADLs) and recreational activities. Both ADLs and recreational tasks have been recognised in the International Classification of Functioning (ICF) activities and participation model by the World Health Organisation (WHO) as integral in gauging the extent of functional limitation experienced by clinical populations (Cieza et al, 2004; WHO, 2001), and have subsequently been identified as important in the assessment of those with LBP (Cieza et al, 2004; Stier-Jarmer, Cieza, Borchers & Stucki, 2009). Despite this point, recreational tasks, particularly those pertaining to sport and social activities, are excluded from the most commonly employed LBP outcome measures, including the Roland Morris Questionnaire (RMQ) (Pengel, Refshauge & Maher, 2004; Sigl et al, 2006).

As physical function is a highly individualised construct, the use of two separate domains in the LLTQ including social and recreational activity may provide a more composite evaluation for different individuals and populations. In doing so, it may also offer a strategy that is more sensitive across specific clinical groups, such as athletes or the elderly, or those with specific occupational demands, and offer a more accurate assessment over various stages of rehabilitation (Spenkelink, Hutten, Hermens & Greitemann, 2002). An additional advantage of the LLTQ is the inclusion of a scoring system for the individual to rate the importance of each task. This allows the clinician or researcher to establish the key requirements and priorities of the individual or clinical group, and furthermore, has the potential to facilitate the weighting of particular functional tasks. This has been identified as an area that may assist in the development of interventions that reflect the priorities of the individual (Higginson & Carr, 2001; McNair et al, 2007). Finally, the LLTQ employs a five point likert system for rating the ability to achieve tasks, as opposed to a dichotomous scale, therefore is able to take into consideration the partial achievement of tasks enabling the detection of smaller yet potentially important improvements and deteriorations in function (Clark & Watson, 1995).

The advantages offered by the LLTQ and furthermore its potential to apply the measure to both those with lower limb conditions and LBP may address some of the issues related to clinician uptake of physical functioning outcome measures by reducing the storage, administration, training and scoring required with multiple tools (Greenhalgh,

Long & Flynn, 2005). The use of one physical function outcome measure that reflects the key priorities of multiple clinical populations whilst remaining reliable, valid and sensitive to important change within these specific populations would be most beneficial to clinicians to maximise practicality. Within research settings, a single tool for evaluating function in those with lumbar and lower limb conditions would potentially assist pooling of data, achieving larger sample sizes and power for research trials (Jaeschke, Singer & Guyatt, 1989).

In order for physical function outcome measures to be utilised, they must be able to accurately reflect physical function as a construct within the population of interest and reflect clinically important change in this construct over time. To evaluate the use of the LLTQ within an acute LBP population, examining the content and construct validity of this measure and establishing the responsiveness and MCID of the measure within this population will assist in determining its value in this setting. As the RMQ is a frequently cited patient reported outcome measure used within this population and has sound psychometric properties (Bombardier, 2000; Rocchi et al, 2005; Roland & Fairbank, 2000; Stratford, Binkley & Riddle, 2000; Turner et al, 2003), it is an appropriate choice as a comparative measure for the LLTQ. If the responsiveness and MCID of the LLTQ were found to be equal to, or better than that of the RMQ whilst demonstrating good validity, clinical application or further investigation of this outcome measure in LBP populations may be warranted.

1.2 Purpose Statement

The purpose of the current study was to investigate the key psychometric properties of the LLTQ in an acute LBP population and compare the findings to that of the RMQ when applied to the same subject group.

The following properties were assessed in an acute LBP sample:

Investigation A; the content validity of the LLTQ through contrast with the ICF core sets for LBP model and analysis of importance ratings of LLTQ tasks.

Investigation B; the construct validity of the LLTQ through correlation analysis of LLTQ scores with those of the RMQ

Investigation C; the statistical responsiveness of the LLTQ using change data and distribution based statistical analyses of effect size and standardised response mean (SRM)

Investigation D; the measurement of error of the LLTQ through calculation of the standard error of measurement (SEM) and minimal detectable change (MDC)

Investigation E; the MCID of the LLTQ and analysis of the MCID according to baseline LLTQ scores.

The order of these investigations is not related to their importance but rather to providing a logical flow to the thesis.

1.3 Significance of the problem

The findings of this study will have significance for health professionals, health researchers, guideline groups and funding bodies who employ outcome measures to detect individual and group level change associated with treatment for acute LBP.

This study will provide important information on use of the LLTQ in acute LBP populations relating to the validity of the outcome measure and ability to detect statistical and clinical important change. In doing so, the LLTQ may offer a way of improving uptake of outcome measures within clinical settings and offer a responsive and valid outcome measure for use within research trials.

CHAPTER 2: REVIEW OF THE LITERATURE

2.1 Introduction

The first objective of this review is to investigate LBP and the limitations in function that are associated with this condition. Next, the measurement properties of validity and reliability are discussed, and are focussed upon the RMQ and LLTQ. As the focus of the Thesis, particular emphasis is given to the concepts of responsiveness and MCID. A semi-systematic review is then undertaken to investigate the MCID of the RMQ and LLTQ, which includes a review of the current methodological practices employed to evaluate the MCID.

2.2 Aetiology and Classification of Low Back Pain

Low back pain is a term given to a range of symptoms including pain, ache, stiffness or fatigue localised below the costal margin and above the inferior gluteal folds, with or without referred leg pain (van Tulder et al, 2006). Individuals with LBP often present with physical signs such as a loss of range of movement, muscle tightness and spasm, reduced or increased curvature of the spine and signs of psychological distress and anxiety (Linton, 2000; Waddell, Somerville, Henderson & Newton, 1992). It is a condition that is often unable to be validated by an external standard and has many possible aeitiologies occurring across all age, gender and occupational populations (Manchikanti, 2000). Lifetime prevalence rates have been estimated at close to 80% (Walker, Muller & Grant, 2004). The condition is often classified as specific or non-specific, pertaining to the respective presence or lack of presence of a recognisable patho-anatomical mechanism (Waddell, Somerville, Henderson & Newton, 1992). With poor diagnostic validity of investigative measures and poor radiological correlation to symptoms, up to ninety percent of all cases are defined as non-specific (Kent & Keating, 2005; Pengel, Herbert, Maher & Refshauge, 2003). Classification of LBP can also be used according to duration of symptoms. Symptoms of less than six weeks is often defined as acute, subacute between six weeks and three months, and chronic, when symptoms persist for longer than three months (Kovacs et al, 2004; Van Tulder et al, 2004).

It is generally accepted that a single episode of acute LBP has a favourable natural history with respect to symptom reduction and restoration of physical function in the short term (Kocavs et al, 2005). In the majority of cases, patients tend to have rapid improvement with resolution of symptoms within one month and ninety percent of patients will recover completely within six weeks (Carey et al, 1995; Coste et al, 1994; Koes, van Tulder & Thomas, 2006; Pengel, Herbert, Maher & Refshauge, 2003). LBP however is often recurrent in nature, with rates as high as 60-86% within the first year of an acute episode (Hides, Jull & Richardson, 2001; Von Korff, Le Resche & Dworkin, 1993; Walker, 2000), thus it is regarded as a major health and socioeconomic problem associated with high costs of health care utilisation, work absenteeism and disablement (Dionne et al, 1997).

2.3 Concepts of Function and Disability in relation to Low Back Pain

Loss of function is an inherent sequelae of LBP, thus it is widely recognised as an important component of patient assessment (Beattie & Maher, 1997). It is also considered a strong prognostic indicator of variables including return to work (Mannion et al, 2001; Nordin et al, 1997). The WHO describe function as a context specific concept involving a combination of individual and societal perspectives by means of activity and participation, via the ICF model (WHO, 2001). The ICF model is divided into categories to represent body structure and function, and activity and societal participation. Because any or all of these factors may be influenced through injury or disablement, the concepts of this model are valuable in developing patient oriented assessment and management (Cieza & Stucki, 2005; Horner & Larmer, 2006; WHO, 2001). Within the United States, Nagi's 'Disablement Model' has also provided a useful concept used by researchers to model consequences of disease and injury at the level of body systems, the individual and society (Jette, 2006). At the level of the individual, Nagi uses the term 'functional limitation' to represent restrictions in the performance of specific tasks by a person, and the term 'disability' referring to the limitation in performing socially defined roles and tasks expected of an individual within a socio-cultural and physical environment, thereby emphasising the highly individualised concept of both functional limitation and disability (Nagi, 1964). Nagi's disablement model has been widely adopted as a conceptual model of function as it has clear terminology, delineates activity, functional and societal limitations and includes

definitions consistent with the concepts included in the WHO model (Jette, Assmann, Rooks, Harris & Crawford, 1998, Jette, 2006).

There is currently debate as to the relationship between limitation in physical function and other impairments and sequelae of LBP including pain, physiological and psychological outcomes (Mannion et al, 2001). Research suggests the most influential group of factors responsible for variance in functional limitation in those with LBP are a combination of pain, psychological distress, fear avoidance beliefs, back muscle activation, lumbar ROM and gender (Jensen & Karoly, 1992; Waddell et al 1993; Jensen et al, 1994; Crombez, Vlaeyen, Heuts & Lysens, 1999; Jensen et al, 1999; Mannion et al, 2001). Authors have however recognised that the complexity of the relationship between these variables limits the ability to accurately predict functional limitation based on the presence or extent of other factors associated with LBP. It is therefore deemed important that function is assessed as a separate construct within patient management (Bombardier, 2000; Mannion et al, 2001).

2.4 Measurement of physical function for Low Back Pain in Physiotherapy

Physiotherapists routinely assess patients within a clinical environment in order to form diagnoses, determine appropriate management strategies and to record patient change as a result of treatment (Horner & Larmer, 2006). Measurement of outcomes as part of this assessment are considered a fundamental requirement of health care provision. National and international physiotherapy bodies and their respective guidelines emphasise the use of appropriate outcome measures to meet basic standards. Registration requirements with the New Zealand Physiotherapy Board also detail competencies that include the collection and utilisation of measures to inform practice (New Zealand Physiotherapy Board 1999).

With the acknowledgement of the impact of LBP on physical function and improvement of function considered an important goal of treatment, many outcome measures assessing this construct have been developed over the last three decades for use in clinical and research settings (Anagnostis, Gatchel & Maher, 2004; Bombardier, 2000; Cieza et al, 2004). Patient reported functional outcome measures in the form of questionnaires offer a way of obtaining an assessment in a standardised, reproducible manner without constraints of time, space and equipment needed for specific physical

function testing. They also offer a more quantifiable and reliable strategy than interviewing and have been shown to be less susceptible to observer bias and inter-observer variation as other methods of measuring physical function (Beaton, 2000; Beattie & Maher, 1997; Greenhalgh, Long & Flynn, 2006). Changes in such outcome measures are considered important in justifying the continuation or change in patient management as the patients priorities are taken into consideration (Greenhough & Fraser, 1992), with current evidence indicating greater satisfaction with care, improved patient-provider communication and shared decision making with use of such patient reported meaures (Fischer et al, 1999; Marshall, Haywood & Fitzpatrick 2006; McHorney, 2002). It is also suggested that patient reported outcome measures offer the potential of increasing communication between health professionals and provide more accurate prediction of time frames for recovery such as return to sport or work (Greenhalgh, Long & Flynn, 2005; Jette, 1993).

To investigate the current practices regarding the use of outcome measurements, Copeland and colleagues (2008) undertook a cross sectional study of clinicians aiming specifically to evaluate the beliefs and attitudes of New Zealand physiotherapists in relation to their use of various LBP measures. This included the use of three well known outcome measures: the Oswestry Low Back Pain Disability Index (Fairbanks, Davies, Couper & O'Brien, 1980), Quebec Back Pain Disability Scale (Kopec et al, 1995) and the RMQ. This study highlighted a lack of clinician uptake of such patient reported physical function outcome measures, and reported that most assessment procedures take place at an impairment level only. Patient reported outcome measures assessing function in other countries including England, Ireland, Australia and Canada are also not well utilised (Beattie & Maher, 1997; Caulfield & Reilly, 2003; Kay, Myers & Huijbrets, 2001; Kirkness &Korner-Bitensky, 2002; May, 2003; Monk, 2006). Kirkness and Korner-Bitensky (2002) examined the records of 265 patients with LBP in 40 physiotherapy practice settings in Canada and found that only 31% of patient clinical notes recorded the use of a physical function outcome measure at initial assessment. The researchers also found that only 10% of notes at follow up assessment and 6% of notes at discharge, included a physical functioning outcome measure. In addition, two European surveys of rehabilitation health professionals, including physiotherapists reported a pattern across professions, countries and health care settings, also indicating that the use of such measures of function in LBP rehabilitation is not routine (, Abrams et al, 2006, Haigh et al, 2001, Torenbeck et al, 2001).

There are several reported explanations accounting for the poor uptake within clinical practice. Practical difficulties include lack of time, money and human resources needed to collect, analyse and make use of results, in addition to insufficient information-technology support for storing and retrieving data (Greenhalgh and Meadows, 1999, Greenhalgh, Long & Flynn, 2005). Specific barriers that have been identified also include a lack of knowledge particularly in regards to which measures to use, how to administer them and how to record and interpret results (Copeland, Taylor & Dean, 2008, Khorsan, Coulter, Hawk & Choate, 2008).

2.5 Practicality, validity and reliability of outcome measures; RMQ and LLTQ

2.5.1 Introduction

In order for an outcome measure to be incorporated into clinical practice, it must minimise the practical burden on the clinician and patient whilst remaining valid, reliable and responsive to meaningful change for a specific population or setting (Hicks, 1999; Horner & Larmer, 2006; Sigl et al, 2006). This section outlines the key measurement properties of outcome measures, discussed particularly in relation to the RMQ and LLTQ. A head to head comparison of these outcome measures with the 'ICF core sets for LBP' model (Stier-Jarmer, Cieza, Borchers & Stucki, 2009) is made to investigate content validity. Although measurement validity is important, it is not the focus of the current thesis. Therefore the last section of the current chapter centres on the concept of responsiveness and the MCID and includes a systematic review of the current evidence base of the MCID of the LLTQ and RMQ.

2.5.2 Practicality, validity and reliability of outcome measures of function for Low Back Pain

The ease of practical application of an outcome measure is vital in gaining acceptance for use in clinical and research settings (Deyo & Carter, 1992). Outcome measures that are time consuming, require complex scoring algorithms or data analysis and interfere with scheduled treatment time are less likely to be adopted (Kopec, 2000). The measure should include explicit time frames regarding functions or tasks within the last 24 hours for example, and should possess a wide enough scale whilst minimising floor and

ceiling effects to include individuals with varying levels of limitation in physical function (Leclaire, Blier, Fortin & Proulx, 1997; Rocchi et al, 2005; Streiner & Norman, 1995;).

Validity is the term given to the ability of an instrument to measure what it is intended to measure when used in a specific population and setting (Jette, 1993). Content validity corresponds to how well the instrument relates to the specific domain of interest. In the area of limitation in physical function, content validity is established through ensuring the items on the measure reflect limitations experienced by individuals with similar conditions (Domholdt, 2005). This is often gained through comparisons to standardised frameworks of function and may include consultation with expert and patient panels (Grotle, Brox & Vollestad, 2005). Assessment of the content validity of LBP physical function outcome measures are somewhat limited due to a reported difficulty evaluating and defining appropriate functional tasks for inclusion in such measures and debate regarding the use of standardised external frameworks of functioning (Grotle, Brox & Vollestad, 2005, Stier-Jarmer, Cieza, Borchers & Stucki, 2009). According to the WHO, the ICF serves as one such framework, representing a universally agreed classification system containing the elements to describe function. Several authors advocate the use of linking the outcome measure to the activities and participation components of the ICF model when considering the assessment of limitation of physical function in individuals with LBP, offering a strategy to compare and standardise items for outcome measures (Cieza et al, 2004; de Vet, Terwee & Bouter, 2003; Grotle, Broz & Vollestad, 2005; Sigl et al, 2006; Stier-Jarmer, Cieza, Borchers & Stucki, 2009).

The following table outlines the identified functional tasks from the activities and participant component of the ICF model in order of importance which are considered most relevant in the assessment of function in those with LBP (WHO, 2001; Cieza et al, 2004).

Table 2.1: ICF Activities and Participation categories considered important in LBP populations (adapted from Cieza et al, 2004).

ICF Category Title	ICF code
Maintaining a body position	D415
Lifting and carrying objects	D430
Changing basic body position	D410
Walking	D450
Remunerative employment	D850
Work and employment, other specified and unspecified	D859
Doing housework	D640
Dressing	D540
Handling stress and other psychological demands	D240
Family relationships	D760
Toileting	D530
Acquiring, keeping and terminating a job	D845

It is apparent from the table that several of the identified functions within the model are not considered physical tasks therefore this model may not offer a complete framework specifically pertaining to important limitation in physical function in LBP populations. Stier-Jarmer and colleagues (2009) have elaborated on this work, with the development of a framework relating to specific tasks within the activities and participation categories of the ICF model, deemed important in the assessment of individuals with LBP. This is detailed further in Table 2.2 within section 2.5.5.

Construct validity refers to the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the constructs of interest (de Vet, Terwee & Bouter, 2003). As the actual physical functioning capacity of an individual cannot be measured directly, construct validity of a outcome measure is often based on comparisons of the likeness of the items within the measure to other established measures assessing the same construct (Grotle, Brox & Vollestad, 2005).

Reliability of an outcome measure relates to the extent to which it yields consistent, repeatable and reproducible estimates of what is assumed to be an underlying true score (Domholdt, 2005). Reliability coefficients are calculated to reflect the proportion of the score which is due to the true score as opposed to the contribution of error. The intra-class correlation co-efficient (ICC) is often the preferred measure of agreement, as it offers a comparison of the variance between subjects, raters and between times of administration (Kopec & Esdaile, 1995; Resnik & Dobrzykowski, 2003), however, a Pearson correlation coefficient or Cronbach alpha can be used to represent consistency between scores over testing (Domholdt, 2005; Rocchi et al, 2005).

2.5.3 Practicality, validity and reliability of the RMQ

The RMQ is a commonly employed outcome measure of physical function in LBP populations. Originally it was designed for use in research but has since been recognised as valuable in assessing individuals within clinical settings where it focuses on the ability to achieve 24 specific physical functional tasks (Roland & Morris, 1983). It is derived from the Sickness Impact Profile, a 136 item health status measure covering all aspects of physical and mental function (Gilson et al, 1975). Several comparisons have since been made between the RMQ and Sickness Impact Profile, with the RMQ found to be quicker and easier to apply clinically (Roland & Fairbank, 2000; Stratford et al, 1994), and more practical in research settings using LBP populations (Deyo et al, 1998). The RMQ is also reported to be the most universally adopted of the patient reported outcome measures specific to LBP. Haigh and colleagues undertook a large scale survey to assess outcome measures used in 418 rehabilitation clinics across Europe for different diagnostic groups. The RMQ was found to be the mostly widely employed of the questionnaire based measures, and the third most commonly used assessment tool when treating those with LBP (Haigh et al, 2001).

A recent review by Grotle, Brox and Vollestad (2005) concluded that the RMQ has strong content validity with respect to the functions outlined in the ICF model, including a spread of functional activities across four of the six relevant categories that include both dynamic and static functional tasks (Grotle, Brox & Vollestad, 2005; Kuijer et al, 2005). Other reports however, have suggested that the RMQ may be better suited in the assessment of individuals with higher levels of functional limitation

(Pengel, Refshauge & Maher, 2004; Resnik, & Dobrzykowski, 2003). Through linking the items of the RMQ, Oswestry Disability Index and North American Spine Society Lumbar Spine Outcome Assessment Instrument with the ICF activities and participation component, Sigl and colleagues (2006) reported that the items of the RMQ were centred on the assessment of lower demanding activity as it focuses on tasks relating to ADLs, domestic life and the use of mobility aids whilst excluding items relating to lifting, pulling, pushing or tasks relating to recreation and leisure such as running and rapid direction changes. Items not assessed within the RMQ also include the two major ICF classification sectors of work and employment and community and civic life (Pengel, Refshauge & Maher, 2004; Sigl et al, 2006). These sectors, which include social, recreational and sporting activity, are constructs of physical function which have been identified as not only adversely affected by LBP, but considered important in the assessment of individuals with LBP (Stier-Jarmer, Cieza, Borchers & Stucki, 2009). The exclusion of such tasks may be considered a major limitation of the RMQ (Cieza et al, 2004; Pengel, Refshauge & Maher, 2004; Stier-Jarmer, Cieza, Borchers & Stucki, 2009).

Regarding the construct validity of the measure, the RMQ has demonstrated good correlation of scores when compared to other established measures of physical functioning, including the physical subscales of the SF-36, the Sickness Impact Profile, the Quebec Back Scale and the Oswestry Disability Index (Jensen, Storm, Turner & Romano, 1992; Kopec et al, 1996; Patrick et al, 1995; Roland & Fairbank, 2000; Stratford et al, 1994) and has shown modest correlation with actual physical performance (Simmonds et al, 1998).

Calculations of reliability co-efficients of the RMQ are generally high; ranging from 0.82 to 0.91 (Deyo & Centor, 1986; Johansson & Lindberg, 1998; Roland & Morris, 1983). Reported calculations specifically pertaining to test-retest reliability have ranged from 0.81 over 48 hours (Stratford, Binkley & Riddle, 2000), 0.88 over 1 week (Johansson & Lindberg, 1998), 0.89 over 1-2 weeks (Underwood, Barnett & Vickers, 1999), 0.93 over 2 days to 2 weeks (Jacob, Baras, Zeet & Epstein, 2001) and 0.91 over an undefined periods up to 2 weeks (Kopec et al, 1995).

From a practical stand point, the RMQ has potential limitations. Although the scoring system of the RMQ is simple to use as responses can be added quickly, the dichotomous nature of the scoring may be a shortfall as it fails to assess partial yet clinically relevant achievement of tasks that may be detected in rating based outcome measures such as the 11 point scale of the Patient specific functional scale (Chapman et al, 1997; Pengel, Refshauge & Maher, 2004) or the 4 point scale of the LLTQ (McNair et al, 2007).

2.5.4 Practicality, validity and reliability of the LLTQ

The LLTQ is a 20 item outcome measure designed for patients with various musculoskeletal lower limb conditions using two distinct domains; assessing ADLs and recreational activities. Each item is assessed on a five point scale for subjects to rate the ability to achieve each task from no difficulty (4 points) to unable to do (0 points), giving a total score out of forty. Patients are also asked to rate the importance of each activity on a Likert scale (1-4) with a score of 4 rated as very important and 1 as not important. By asking the patient to rate their difficulty performing the task in the last 24 hours, the measure ensures accurate patient recall, accounting for change in physical function that can occur over short periods of time, often demonstrated in individuals with acute LBP (Kopec et al, 2005).

Although the LLTQ has not been examined in relation to an external framework of function, the authors of the LLTQ developed the 20 items based on a review of literature assessing outcomes in those with lower limb conditions and consultation with an expert panel (McNair et al, 2007). McNair and colleagues (2007), assessed the construct validity of the LLTQ by correlating scores from the measure with other measures of lower limb function and general limitation in physical function. The LLTQ correlated moderately with the Lysholm Knee Rating Scale (Lysholm & Gillquist, 1982), Cincinnati Knee Rating Scale (Barber-Westin, Noyes & McCloskey, 1999), Ankle-Hindfoot Scale (Kitaoka et al, 1994) and PSFS (Chatman et al, 1997), and demonstrated a high correlation with the physical functioning component of the SF-36 (Ware & Sherbourne, 1992).

Test- retest reliability estimates of the LLTQ over 1 to 7 days have revealed high levels of reliability with ICCs of 0.96 and 0.98 for the ADL and Recreational domains respectively (McNair et al, 2007).

2.5.5 Comparison of content of RMQ and LLTQ in relation to the ICF model

As the activity and participation component of the ICF model has been identified as representative of important aspects of an individual's physical functioning (WHO, 2001), the LLTQ and RMQ are contrasted in table 2.2 according to those items of the components deemed most important in the assessment of physical function in those with LBP (Cieza et al, 2004; Stier-Jarmer, Cieza, Borchers & Stucki, 2009).

Although the 24 point RMQ includes four more items than the 20 item LLTQ, the LLTQ covers a wider spread of functional tasks in relation specifically to four classification areas compared to the RMQ, as based on the ICF model. These tasks specifically relate to social and sporting activities, and work and employment. As a result, the LLTQ may potentially offer a more composite evaluation of an individual's physical functioning capacity based on this framework.

Table 2.2: Comparison of Content of the Outcome Measures: Contrast to the ICF Core Sets for LBP Model (Stier-Jarmer, Cieza, Borchers & Stucki, 2009)

ICF Classification	RMQ	LLTQ
D4. Mobility		
Sitting down and getting up	+	+
Bending or stooping	+	+
Lying down and getting up	-	+
Standing	+	+
Sitting	+	-
Reaching	-	-
Pulling/Pushing	-	+
Throwing/Catching	-	-
Lifting/Carrying	-	-
Walking	+	+
Running	-	+
Climbing Stairs	+	+
Moving around using transportation	-	+
D5. Self Care		
Washing and Grooming	-	-
Dressing	+	-
Putting on/off footwear	+	-
D6. Domestic Life		
Housework	+	-
D7. Interpersonal interactions and intimate relationships		
General interpersonal interactions	+	-
Intimate relationships	-	-
D8. Major Life Areas		
Work and Employment	-	+
D9. Community Social and Civic Life		
Social Activities	-	+
Sporting Activities	-	+

+ Included in the measure   - Not included in the measure

2.6 Responsiveness and Minimal clinical important difference of outcome measures

2.6.1 Concepts of Responsiveness and Minimal clinical important difference

Scales designed to measure outcomes in patients over time are expected to possess accuracy in detecting meaningful change for the patient, a concept often referred to as responsiveness (Taylor, Taylor, Foy & Fogg, 2001). Most authors agree that responsiveness involves the ability of a measure to detect change despite the wide variety in opinion regarding the nature of change that is being detected (Beaton, 2000). For example, Guyatt, Kirshner & Jaeschke (1992) define responsiveness as the ability to detect important change in the way patients are feeling, even if those changes are small, whereas Testa and Nackley (1994) define responsiveness as the ability to detect meaningful treatment effects. These definitions however, leave the constructs of 'feeling' and 'treatment' open to various interpretation. De Bruin and colleagues rather, explain responsiveness as the ability of an instrument to accurately detect change when change has occurred (De Bruin et al, 1997). This definition has been adopted by other authors to represent responsiveness as it requires determination that change is demonstrated, but does not specify the nature of the change (Beaton, Bombardier, Katz & Wright, 2001).

Responsiveness as a measurement property is a highly context specific attribute. It is determined by the nature of the outcome measure itself, the clinical setting in which it is applied, the patient population or individual being measured, time periods used between testing and the type of change anticipated or observed (Beaton, Bombardier, Katz & Wright, 2001). As a result, clinicians and researchers are required to not only consider the evidence for the responsiveness of the measure itself, but how the setting in which the outcome is applied may influence levels of change (Beaton, Bombardier, Katz & Wright, 2001; Farrar et al, 2000; Fischer et al, 1999). Beaton and colleagues (2001) have addressed the concepts of evaluating responsiveness by describing a constructive 'taxonomy' to describe the considerations that must be given when measuring change. The authors propose that the use of a taxonomy reconciles much of the debate by locating the nature of change within a matrix of three axes; Who is the focus of the study? When were the two measures that are being compared gathered? And, what is the nature of change being examined? These axes define the essential attributes of all

types of change that have been or may be studied under the concept of 'responsiveness' providing clinicians and researchers with a framework of reflection that should be considered when evaluating change (Beaton, Bombardier, Katz & Wright, 2001; Wells et al, 2001).

The concept of the MCID was first defined by Jaeschke and colleagues, as the smallest difference in score in the measure of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive costs, a change in the patients management (Jaeschke, Singer & Guyatt, 1989). The MCID, by incorporating what the patient deems as important change therefore includes consideration of the validity of change scores (Bombardier, Hayden & Beaton, 2001, de Vet et al, 2006).

Some authors contend that the MCID is the most important measurement property to consider when considering clinical decisions relevant to individual patients, as it allows the clinician or researcher to review scores over time and determine whether important change for the patient is likely to have been achieved (Beaton, Bombardier, Katz & Wright, 2001; Childs, Piva & Wright, 2005; Crosby, Kolotkin & Williams, 2004). As a result, the MCID allows clinicians and researchers to take into account the individuals perception of change and therefore adopt changes based on the patients perspective (Ostelo & de Vet, 2005; Revicki et al, 2006). Knowledge of the MCID of an outcome measure may also assist in the calculation of sample sizes and when estimating the clinical relevance of outcomes in studies regarding the efficacy of interventions (Jaeschke, Singer & Guyatt, 1989; Terwee et al, 2003).

Despite these points, there is wide debate within the literature as to the most accurate and appropriate method of establishing the MCID of outcome measures. The lack of standardisation in methodological approaches has been identified as a potential reason, or at least a contribution, to the large discrepancies in MCIDs generated for individual outcome measures (Bombardier, Hayden & Beaton, 2001; Terwee et al, 2010). Notwithstanding, it is widely accepted that identifying a clinically meaningful change requires some form of anchor to define important change based on the individuals perception of meaningful change, in addition to distribution based approaches to estimate statistical responsiveness (de Vet et al, 2007; Guyatt et al, 2002; Leidy & Wyrwich, 2004, Revicki et al, 2006, Wyrwich et al , 2007).

2.6.2 Distribution based approaches

Distribution based methods of assessing responsiveness of a measure differ from anchor based methods in that they provide a means for establishing change beyond some level of random variation (Crosby, Kolotkin & Williams, 2003). Distribution based indices include assessing the magnitude of change through the effect size and SRM using the mean score changes and variation of the population at study (de Vet et al, 2006). The SEM which employs an estimate of outcome measure reliability and population variation, and the MDC using a chosen confidence level, offer estimates of error in the same units as the outcome measure under investigation (Revicki, Hays, Cella & Sloan, 2008).

Because distribution based methods take into account the variation of the sample to calculate change, they do not rely on the use of an external measure and are therefore not affected by issues relating to the validity and biases of such measures (Guyatt et al, 2002; Norman, Stratford & Regehr, 1997). Another advantage of this approach is that measures of variability are always available and therefore values are easy to generate providing a simpler interpretation of results (Cella et al, 2002; Kazis, Anderson & Meenan, 1989; Lydick & Epstein, 1993). The major limitation of distribution based approaches however, is that without reference to an external criterion of change, they are only able to provide an estimation of statistically relevant change, not change that is deemed meaningful to the patient or population at study (Crosby, Kolotkin & Williams, 2003, Copay et al, 2007). Another limitation of this approach is the influence of the variability of the population at study. If the population enrolled is highly heterogeneous with a large variability in scores, the distribution based calculation will be affected, and the score will require a larger degree of change to meet statistical significance. The converse is also true however, where a very homogeneous population could lead to an underestimation of the amount of change required to be statistically relevant (Crosby, Kolotkin & Williams, 2003; Farrar et al, 2000, Guyatt et al, 2002).

The following section provides further details of the key distribution based approaches commonly used in evaluating the responsiveness of outcome measures.

<u>2.6.2.1 Effect size and Standardised response mean</u>

The effect size offers a quantitative estimate of the difference between groups of subjects or groups of scores, therefore can be used to provide an index of statistical responsiveness for different outcome measures using the variability of baseline scores (Copay et al, 2007). It is calculated by dividing the difference in outcome measure change scores from baseline to follow up by the standard deviation of baseline scores (Copay et al, 2007). The advantage of effect size statistics is that they are able to translate the magnitude of change into a standard unit of measurement that facilitates a comparison among various outcome measures within a given subject population and setting (Hevey & McGee, 1998). Cohen describes effect sizes as small (0.2) moderate (0.5) and large (0.8), representing the number of standard deviations by which the scores change over the time of testing (Cohen, 1977).

Critics of the effect size as a measure of statistical responsiveness argue that the calculation should be derived from variance in change scores, rather than variance in baseline scores (Diehr et al, 2005; Katz, Larson, Phillips, Fossel & Liang, 1992). This has led to the development of the SRM which uses the standard deviation of change scores as the denominator to represent responsiveness of the measure. The SRM represents the signal to noise ratio and is defined as the ratio of mean change to the standard deviation of the change scores in the population of patients reporting change (Hurst et al, 1997; Katz et al, 1992). Similar to effect size calculations, the SRM offers a comparable statistic that can be contrasted across various outcome measures.

<u>2.6.2.2 Standard error of measurement and Minimal detectable change</u>

The SEM and MDC offer measures of error representing the change in score that is required to be considered statistically reliable (Beaton, Boers & Wells, 2002; de Vet et al, 2006; Ferguson, Robinson & Splaine, 2002; Stratford et al, 1996). As the MDC calculation employs a level of confidence within the statistic, it assumes that there is only a small chance, represented by this confidence level, that the patients change is contributable to error alone when the score change is greater than the MDC (Resnik & Dobrzykowski, 2005).

The SEM uses a reliability parameter of the given outcome measure and the variability of subject scores given by the standard deviation, to provide an estimate of the ability of the outcome measure to detect change that is above that due to error alone. The SEM is

calculated as the standard deviation*√ (1-R). The value used to represent the estimate of reliability varies, and is often represented by a Pearson product co-efficient, ICC and in some cases, the Cronbach alpha (Crosby, Kolotkin & Williams, 2003).

Several authors (Crosby, Kolotkin & Williams, 2003; Roebroeck, Harlaar & Lankhorst, 1993; Wyrwich, Tierney & Wolinsky, 1999) have argued that the SEM is the most appropriate calculation for determining statistically meaningful change based on several particular properties. The SEM accounts for the contribution of random error to observed change in the measure and is considered to be a fixed characteristic of the measure, as the standard deviation and the reliability coefficient remain relatively constant across samples taken from a given population. The SEM does however make the assumption that measurement error is stable across possible subject scores. This is a notable limitation as it has been established that error is highly dependent on where the given score falls on the scale with those falling nearer the centre of a scale having higher levels of error than those found at the ends of the scale, where less variability is encountered (Binkley & Stratford, 1999; Stratford, Binkley & Solomon, 1996). In addition, variation in the statistic used as the reliability co-efficient may account for differences in the SEM values generated (Crosby, Kolotkin & Williams, 2003; Wyrwich, Nienaber, Tierney & Wolinsky, 1999).

The MDC is a term given to the smallest change in score that can be considered above measurement error, with a given level of confidence so that a score change below this value is deemed to be indistinguishable from measurement error (Beaton, Boers & Wells, 2002, de Vet et al, 2006). The statistic can be set at a given confidence level, often 90% or 95%, and employs the square root of two within the calculation to adjust for the error associated with taking two measurements (Hebert, Spiegelhalter & Brayne, 1997; Nunnally & Bernstein, 1994). It is generated in the same units as the outcome measure and therefore can be directly contrasted to the MCID (Ferguson, Robinson & Splaine, 2002; Stratford et al, 1996). As the SEM is included within the calculation, the statistic is influenced by the sample variation and the chosen reliability parameter of the given measure (Beaton, Boers & Wells, 2002). The calculation of the MDC, considered to reduce the effect of 'background noise' associated with the study, will therefore become larger with increased sample variation and lower parameters of reliability. Several authors have recognised this limitation (de Vet et al, 2006; Lassere et al, 2001; Stratford, Binkley et al, 1996), where the statistic has found to be particularly

susceptible to sample variation and low reliability estimates, often resulting in large, conservative MDC's (de Vet et al, 2006, Wyrwich, Nienaber, Tierney & Wolinsky, 1999, Wyrwich, Tierney & Wolinsky, 1999). As a result, in practice, subjects may have to achieve high changes in the target outcome measure to reach a score change that is above the calculated MDC (de Vet et al, 2006, van der Roer et al, 2006).

2.6.3 Anchor based approaches and the Global perceived effect score

Anchor based approaches are widely recommended for establishing the MCID of outcome measures as they consider meaningful change from the subjects perspective thereby increasing the relevance of the statistic in clinical settings (Beaton, Bombardier, Katz & Wright, 2001, Beaton, Boers & Wells, 2002; Guyatt et al, 2002; Stratford et al, 1994).

In an anchor based approach, patient global ratings of change are considered to be especially well suited to assess important change from the patients perspective, however this rating system must be valid, reliable and practical in a clinical setting (Beaton, Boers & Wells, 2002; Guyatt et al, 2002). Research indicates that global change assessments as anchors are more sensitive to change than other outcome measures, have adequate psychometric properties, are strongly correlated with patient priorities and perspectives and have the potential to take into account more information that may affect an individuals function (Farrar et al, 2000; Hagg, Fritzell, Oden & Nordwall 2002; Revicki et al, 2006). It is therefore argued that global change measures should be used to gauge overall change, giving a comprehensive evaluation over time and interventions. (Crosby, Kolotkin & Williams, 2004; de Vet et al, 2006; Farrar, et al, 2001; Fischer, Fritzell, Oden & Nordwall, 1999).

The Global Perceived Effect Score (GPE), also commonly referred to as the patient global impression of change, is a self rated instrument often employed as the anchor within research to determine important change. There is considerable variation in the design and structure of the GPE, particularly relating to the number of items on the scale and labels or terminology used to indicate levels of change (Dworkin et al, 2005; Kamper, Maher & Mackay, 2009; Ostelo & de Vet, 2005). The level on the scale which is deemed a definition for minimal meaningful change may vary, however most authors define this level as 'better' or 'improved', depending on terminology used, and a level

of 'slightly improved' or less to be indicative of no clinical improvement (Beuskens, de Vet & Koke, 1996, Stratford, Binkley, Riddle & Guyatt 1998; Farrar et al, 2001; Ostelo, de Vet, Knol & van den Brandt, 2004; Dworkin et al, 2005; Ostelo & de Vet, 2005).

Norman and colleagues (1997) raise three particular concerns regarding the use of GPE scales as an external criterion to determine meaningful change. The authors argue that the reliability and validity of such measures are not fully established, they are highly correlated with the individuals present status and bias in the individuals judgment of change will also be reflected in the final outcome measure score (Norman, Stratford & Regehr, 1997). The potential error encountered with the use of the GPE to assess overall change by relying on the ability of the patient to recall previous functional ability is a concern reported widely within the literature (Copay et al, 2007; Redelmeier, Guyatt & Goldstein, 1996; Norman, 2003). Bias may result from patients recalling more pain or limitation in physical function previously than actually reported at the time (Norman, Stratford & Regehr, 1997). Potential explanations for this phenomenon include Ross' implicit theory of change (Ross, 1989). This theory suggests that when a patient is faced with the challenge of estimating change, they start with their current state and move backward rather than accurately recall their previous state. In addition, patients with obsequious behaviour may systematically alter responses in the direction that they perceive the clinician desires (Norman, 2003). Another potential limitation when relying on an external measure as an anchor is the completion of the investigated outcome measure concurrently with the anchor, as errors encountered in both measures will likely be correlated. This may lead to similar biases occurring during measurement administration (Kamper, Maher & Mackay, 2009; Norman, 2003; Redelmeier, Guyatt & Goldstein, 1996; Stratford, Binkley, Riddle & Guyatt, 1998).

Despite reservations in using such scales, the measurement properties of the GPE have been investigated and are widely considered the most appropriate tool for determining meaningful change within clinical populations, including LBP (Copay et al, 2007; Dworkin et al, 2005; Farrar et al, 2000; Hagg, Fritzell & Nordwall, 2003; Ostelo & de Vet, 2005). A major strength of the GPE is the high level of face validity of the measure. Fischer and colleagues (1999) investigated the correlation between GPE scales and patient satisfaction measures, with results showing correlations significantly higher than non-retrospective measures. Several authors have also found the GPE to be more

sensitive to change compared to serial measurements (Aseltine, Carlson, Fowler & Barry, 1995; Fischer et al, 1999; Lauridsen et al, 2006), particularly in acute clinical populations who are at less risk of recall and motivational biases (Lauridsen et al, 2006). The 11 point scale has also shown high test-retest reliability (ICC = 0.90, Costa et al, 2008), and high statistical responsiveness based on SRM measures have been demonstrated using the 7 point and 15 point scale (Fischer et al, 1999; Lauridsen et al, 2006). Construct validation has included correlation analysis with limitation in physical function with numerous studies demonstrating significant correlation between GPE scores and change in physical function outcome measures (Costa et al, 2008; Kamper, Maher & Mackay, 2009; Kopec & Esdaile, 1995; Little & McDonald, 1994; Stratford et al, 1994). There is currently no consensus within the literature as to the number of items and terminology used in the design of the scale, nor is there agreement as to the influence of variation in the use of GPE scales on the resultant MCID (Demoulin, Ostelo, Knottnerus & Smeets, 2010; Kamper, Maher & Mackay, 2009; Revicki, Hays, Cella & Sloan, 2008; Terwee et al, 2010).

Analytical strategies when using an anchor based approach to determine the MCID involve the use of receiver operator characteristic (ROC) curves to establish the cut off point where a particular change in score will be considered optimal to represent meaningful change. The ROC curve generated from the results plots the 'true positive rate' (sensitivity) against the 'false positive rate' (1-specificity) for various change score 'cut points'. The cut point is the change score above which the result will be a meaningful improvement, and below will be no improvement as defined by the anchor. The resulting area under the generated curve indicates the probability of making a correct ranking of an improved or non-improved subject, according to their change in the outcome measure of interest. A value of 1 under the area of the curve would indicate 100% identification of a meaningful improvement, while 50% would represent a decision that is no better than chance alone. An optimal outcome to indicate a highly responsive measure would be where both a high sensitivity and specificity were generated (Guyatt et al, 2002, Revicki, Hays, Cella & Sloan, 2008, Ward, Marx & Barry, 2000).

Potential difficulties and limitations encountered with using these methods of analysis to generate the MCID include the reliance on the validity and reliability of the anchor as a representation of meaningful change and the selection of appropriate cut off points to distinguish those who have experienced meaningful change or not (Copay et al, 2007; Norman, 2003; Terwee et al, 2010). In addition, the method requires the dichotomous grouping of individuals according to those who have experienced at least the defined level of minimal improvement on the anchor and those who have not, thereby failing to take into consideration the amount of improvement or deterioration experienced. The advantages however of an anchor based approach is that the MCID is generated in the same point system as the outcome measure, providing a simple comparative value to apply in clinical and research settings (Copay et al, 2007; Stratford, Binkley & Riddle, 1996). In addition, the chosen anchor offers a single measure of change from the individuals perspective aggregating all of the components of the subjects experience into one overall measure of change (Crosby, Kolotkin & Williams, 2003).

2.6.4 Combining methodological approaches to generate the Minimal clinical important difference

Over the past several years, there has been recommendations made for methodological practices which combine various distribution and anchor based approaches to establish a single MCID for an outcome measure (Revicki, Hays, Cella & Sloan, 2008). Often referred to as 'triangulation' these approaches are used on the premise that the true MCID of an outcome measure lies somewhere between the values generated using anchor and distribution based methods (Revicki et al, 2006; Revicki, Hays, Cella & Sloan, 2008; Terwee et al, 2010). It is also suggested that using multiple approaches may act to increase reliability and reduce systematic error (Leidy & Wywich, 2004), although the value of such approaches have been largely based on research addressing the use of health-related quality of life measures (Guyatt et al, 2002; Revicki, Hays, Cella & Sloan, 2008; Wyrwich et al, 2005)

Revicki and colleagues (2008) have suggested that an estimation of an MCID be based on multiple approaches including a triangulation of statistical methods with more weighting being placed on anchor based methods than distribution based data as the latter does not provide any direct evidence of meaningful change. The authors also suggest using review methods and modified Delphi methods to assist in the selection of

an appropriate MCID for an outcome measure. The shortfalls in generating the MCID for a measure using multiple approaches however have been highlighted recently by Terwee et al (2010). The researchers applied several distribution and anchor based approaches to the change scores of physical functioning outcome measures in hip and knee populations receiving treatment in various settings. The results demonstrated significant variations in MCIDs, despite using the same analytical methods across different studies and different methods within the same studies.

Overall, there is little consensus within the literature as to the most appropriate and accurate methodological practice to estimate the MCID using either single or multiple analytical approaches (Demoulin, Ostelo, Knottnerus & Smeets, 2010; Revicki et al, 2006; Revicki, Hays, Cella & Sloan, 2008; Terwee et al, 2010). As a result, recent reports have highlighted the caution needed when interpreting published MCIDs for clinical use (Demoulin, Ostelo, Knottnerus & Smeets, 2010; Terwee et al, 2010). Terwee and colleagues (2010) have recently suggested that due to such discrepancies in methods undertaken to investigate the MCID of outcome measures, the use of MCIDs based on published research must take into consideration the methodologies of the individual studies. In light of this, a review of the literature of research pertaining to the MCID of the RMQ and LLTQ is undertaken to assess the methodological practices of the studies and potentially ascertain the MCIDs of these outcome measures.

2.7 Minimal clinical important difference of the RMQ and LLTQ; a Review of the Literature

2.7.1 Introduction

This review outlines the current evidence base for the MCID of the RMQ and LLTQ. It follows the normal processes of a systematic review, however only one researcher has carried out assessment of the selected papers. The methodological rigour of this evidence based is assessed using a quality evaluative tool with key areas identified from this research base discussed further.

2.7.2 Review Methods

<u>2.7.2.1 Search Strategy</u>

A search for studies investigating the MCID of the RMQ and LLTQ were performed between February and June 2009.

Initially a variety of sources were utilised including national and international journals and clinical and research guidelines, in addition to a general internet search. From this initial search, a keyword list was developed to encompass definitions of responsiveness and MCID associated with the LLTQ and RMQ. The search terms are outlined in Table 2.3.

Table 2.3: Search terms: Keywords used in search

Keywords

Anchor	Minimal important difference
Important difference	LLTQ
Meaningful change	Low back pain
MCID	Lower limb tasks questionnaire
MDC	Responsive(ness)
MIC	RMQ
MID	RMQ-24
Minimal clinical difference	ROC curve (analysis)
Minimal clinically important difference	Roland Morris (questionnaire)
Minimal detectable change	Sensitivity to change
Minimal important change	Smallest detectable change

An initial check of the keyword list was made against each of the subject headings from 11 electronic databases; Allied and Complementary Medicine (AMED, 1994+), Auckland University of Technology Library Catalogue including E-Journals, Cochrane Library, Cumulative Index to Nursing & Allied Health Literature (CINAHL, 1994+), Google, Google Scholar, ISI Web of Knowledge: Current Contents Connect (1994+), EBSCO Megafile Health Premier (including Medline, Health Source: Consumer Edition and Nursing/Academic Edition), Medline via PubMed, PEDro (Physiotherapy

Evidence Database) and Sports Discus. Search terms were modified as required for each database and were used individually and combined as phrases using Boolean operators to target studies in each category. The search terms were used in default fields and in specific fields such as abstract, journal and title as required. The names of key authors were also identified from the initial search and used as search terms within the author fields of the databases. The literature search was supplemented with a review of the bibliographies of past review papers.

2.7.2.2 Study selection

Studies of the MCID of the RMQ or LLTQ were identified for inclusion, and were required to meet the additional criteria;

- Published within the past 15 years (Jan 1994–May 2009).
- English language publications
- Anchor based methods of calculation used to generate the MCID
- Not published in the popular press, such as magazines and newspapers.

Full copies of all included studies were obtained. Due to constraints on time and resources, one researcher reviewed each included article. A second reviewer was consulted where study eligibility was questioned.

2.7.2.3 Quality assessment; The modified QUADAS tool

Assessing the quality of methodology and reporting of studies including susceptibility to bias is essential in the interpretation of an evidence base when conducting a review (Juni, Altman & Egger, 2001). There is currently no specific tool for evaluating research pertaining to the responsiveness or MCIDs of outcome measures. For the purpose of the review, a modified version of the QUADAS tool was formulated to highlight the methodological quality of the selected papers. No formal scoring was used with this tool as it was employed only to give an appreciation of the methodological rigour of the evidence base.

The original QUADAS tool (Whiting et al, 2003) has been specifically designed for assessing the quality of diagnostic accuracy studies, where its role is to gauge overall bias and validity of studies assessing health outcome measures (Whiting, Harbord & Kleijnen, 2005). It consists of 14 items phased as questions, scored as 'yes', 'no' or 'unclear', with operational standards developed for each item (Whiting et al, 2003). The

QUADAS tool has been validated and has demonstrated good agreement between reviewers and across individual items within the checklist (Whiting et al, 2006, Cook, Cleland & Huijbrets, 2007, Meads & Davenport, 2009). Meads & Davenport, (2009) have recommended the QUADAS as a starting point to evaluate the evidence base for studies assessing the use of outcome measures, whilst other authors (Hollingworth et al, 2006) have recommended the use of the QUADAS for highlighting the strengths and weaknesses of existing studies.

In the modification of the QUADAS tool (Table 2.4), several items have been altered from the original version in order to focus on quality issues relevant to studies of the MCID of outcome measures. Modified items and items removed from the original tool were made by consensus between the reviewers. Table 2.4 outlines the items of the modified QUADAS tool.

Table 2.4: Modified QUADAS Scale

Item

1   Were subjects representative of those who will receive the outcome measure in practice?

2   Were selection criteria clearly described?

3   Was a patient reported anchor employed?

4   Were the anchor and outcome measure completed on the same occasion?

5   Was the completion of the outcome measure and anchor described in sufficient detail to permit its replication?

6   Was the SEM and/or MDC calculated for contrast with the MCID?

7   Was a distribution based method used in the calculation of responsiveness?

8   Was there sufficient reporting of the results of the study?

9   Was the MCID generated according to initial scores of the outcome measure?

10   Were withdrawals from the study explained?

11   Were likelihood ratios and/or percentage changes developed for the results?

12   Were patients blinded to their initial scores at follow up?

2.7.3 Results of Review

2.7.3.1 Literature search results

From the initial literature search, 57 articles were identified and deemed suitable for review of abstracts.

At this stage of the review, 41 studies were excluded. The primary reasons for exclusion included the RMQ or LLTQ not being investigated as an outcome measure and studies not employing anchor based methods within the methodology, an integral component in generating the MCID of an outcome measure (Copay et al, 2007; Revicki, Hays, Cella & Sloan, 2008; Stratford, Binkley & Riddle 1996; Terwee, 2010). Nineteen studies received a full article review, of which nine were then found to be suitable for appraisal. The main reason for exclusion of the remaining studies was the lack of an anchor based approach to generate the MCID of the outcome measure. These remaining ten articles and reasons for their exclusion from the review are located in Appendix 2.

The nine remaining articles, of which eight investigated the MCID of the RMQ and one investigated the LLTQ, were then subject to appraisal and scoring.

2.7.3.2 Data extraction and synthesis

One researcher extracted data from the included studies. The retrieved articles were coded and recorded in Endnote 2007. Data were tabulated under the headings; subjects, outcome measures, testing intervals, definition of meaningful change and results including the MCID according to initial scores (Table 2.5). Where possible, anchor and distribution approach based data were extracted. Any other results reported in the studies were also recorded.

Table 2.5: Studies Selected for Appraisal

Author	Subjects	Outcome Measure	Testing Intervals	Definition of meaningful change	Results (MCID)
Beurskens, de Vet & Koke, 1996	N = 81, non-specific LBP > 6/52 duration	24 pt English RMQ	Baseline and following 5/52 of physiotherapy	'Improved' using 7 Point GPE	Optimal cut off value of 2.5 to 5 points
Stratford, Binkley, Riddle & Guyatt, 1998	N = 226 acute non specific LBP < 6/52 duration	24 pt English RMQ	Baseline and following 3-6/52 of physiotherapy	'Improved' using Patient and therapist 15 point GPE	2, 4, 5, 8 and 8 points, for initial score intervals of 0-8, 5-12, 9-16, 13-20 and 17-24 points respectively
Ostelo, de Vet, Knol & van den Brandt, 2004	N = 105, those with pain at 6/52 post lumbar disc surgery	24 pt English RMQ	Baseline (6/52 post surgery) and at 12/52 later	'Improved' using 7 point GPE	Optimal cut off point of 3.5 points
Grotle, Brox & Vøllestad, 2004	N = 54 acute (<3/52) and 50 chronic (>3/52) LBP patients	24 pt Norwegian RMQ	Baseline and following 4/52 (acute group) or 3/12 (chronic)	'Improved' using 6 point GPE	Optimal cut off point of 2.5 RMQ points for acute group or 1.5 in chronic group
Davidson and Keating, 2005	N = 106, non specific LBP any duration	24 pt English RMQ	Baseline and following 6/52 of physiotherapy	'Improved' using 7 point GPE	MCID of 8-9 RMQ score points.

Table 2.5 Continued: Studies Selected for Appraisal

Author	Subjects	Outcome Measure	Testing Intervals	Definition of meaningful change	Results (MCID)
Jordan, Dunn, Lewis & Croft, 2006	N = 447, non specific LBP any duration	24 pt English RMQ	Baseline and at 6/12 following physiotherapy	'Improved' using 6 point GPE	Score reduction of 30%
Lauridsen et al, 2006	N = 233, non specific LBP with and without leg pain any duration	23 pt Danish RMQ	Baseline, 1/52 and 8/52 following treatment.	'Improved' using 7 point GPE and at least 7 points on a 10 point numerical rating scale of 'importance of improvement'	Score reduction of 38%, absolute score given as 5 RMQ points, dependent on baseline.
Kovacs et al, 2007	N = 1349 subacute (<6/52) and chronic (>3/12) LBP with or without leg pain	24 pt Spanish RMQ	Baseline and 12/52 following treatment	'Improved' using 4 point GPE	3.5 – 12.1 RMQ points, dependent on baseline score
McNair et al, 2007	N = 119, various lower limb conditions	LLTQ	Baseline and at 7-10 days.	'Better' using patient and therapist 5 point GPE	4 points in both domains (using mean of anchor, distribution and likelihood ratio approach).

<u>2.7.3.3 Quality assessment</u>

All nine studies critically appraised in the current review employed anchor based methods using a patient reported GPE scale as the external criterion of change. All studies employed the GPE at follow up at the same time as the outcome measure assessment and reported such procedures within the methodology. One study did not include analysis of responsiveness using distribution based statistical methods (Stratford, Biknley, Riddle & Guyatt, 1998). No reviewed studies made follow up assessments with subject withdrawals. Only Stratford and colleagues (1998) were clear as to the blinding of patients to their initial outcome measure scores. Four studies (Jordan, Dunn, Lewis & Croft, 2006; Kovacs et al, 2007; Lauridsen et al, 2006; Stratford, Binkley, Riddle & Guyatt, 1998) made additional calculations of the MCID based on initial or baseline outcome measure scores.

The results related to the quality assessment using the modified QUADAS tool are outlined in Table 2.6. As the modified QUADAS tool was employed to give an appreciation of the methodological quality of the studies, scores were not added.

Table 2.6: Quality Assessment of Studies using Modified QUADAS Scale

Item	Beurskens, de Vet & Koke 1996	Stratford, Binkley, Riddle & Guyatt, 1998	Ostelo, de Vet, Knol & van den Brandt, 2004	Grotle, Brox & Vøllestad, 2004	Davidson and Keating, 2005	Jordan, Dunn, Lewis & Croft, 2006	Lauridsen et al, 2006	Kovacs et al, 2007	McNair et al, 2007
1	+	+	-	+	+	+	+	+	+
2	+	+	+	+	+	-	+	+	+
3	+	+	+	+	+	+	+	+	+
4	+	+	+	+	+	+	+	+	+
5	+	+	+	+	+	+	+	+	+
6	-	-	+	-	+	+	-	+	+
7	+	-	+	+	+	+	+	+	+
8	+	+	+	+	+	+	+	+	+
9	-	+	-	-	-	+	+	+	-
10	-	-	-	-	-	-	-	-	-
11	-	+	-	-	-	+	+	-	+
12	?	+	?	?	?	?	?	?	?

+Achieved   - Not Achieved   ? Unclear

<u>2.7.3.4 Subjects</u>

Across the nine studies, 2762 subjects were involved. Within and across studies investigating the MCID of the RMQ, the number of subjects participating ranged from 50 (Grotle, Brox & Vøllestad, 2004) to 1349 (Kovacs et al, 2007). The gender ratio's varied from 41% female (Ostelo, de Vet, Knol & van den Brandt, 2004) to 73% (Grotle, Brox & Vøllestad, 2004), and the mean age of subjects ranged from 38 (Grotle, Brox & Vøllestad, 2004) to 54 (Kovacs et al, 2007). One study did not report the mean age of subjects (Davidson et al, 2002). Subjects were recruited from a variety of settings including hospital outpatient services, community health centres and private practices and in one study (Ostelo, de Vet, Knol & van den Brandt, 2004), recruitment was solely through post surgical follow up consultation.

Of the eight studies assessing the MCID of the RMQ, the clinical population according to duration of symptoms varied widely. Symptoms of longer than 6 weeks was a requirement for participation within one study (Beurskens, de Vet & Koke, et al, 1996), whereas Stratford and colleagues (1998) recruited only those with acute LBP of less than 6 weeks duration. Grotle and colleagues (2004) recruited two separate cohorts; acute LBP of less than 3 weeks duration and those with chronic LBP, defined as symptoms of longer than 3 months. Kovacs and colleagues (2007) also differentiated subjects according to symptom duration, recruiting subjects with only subacute and chronic LBP, defined as more than 14 days and 90 days respectively. Ostelo and colleagues (2004) recruited only those subjects with persisting LBP at a 6 week review following lumbar disc surgery, and the three remaining studies (Davidson & Keating, 2002; Jordan, Dunn, Lewis & Croft, 2005; Lauridsen et al, 2006) recruited those with non-specific LBP of any duration. The eight RMQ studies had similar inclusion criteria, where subjects were required to be aged over 18 and have the presence of LBP without the presence of a pathological disorder of the spine including fractures, infection or malignancy. The remaining study included within the review (McNair et al, 2007) investigated the MCID of the LLTQ. Data from 119 subjects with a mean age of 34 were recruited from outpatient settings with a variety of acute and chronic lower limb injuries, where a rehabilitation period of up to 6 weeks was anticipated.

<u>2.7.3.5 Outcome measures and testing points</u>

The primary outcome measure investigated in five of the reviewed studies (Beurskens, de Vet & Koke, 1996; Davidson & Keating, 2002; Jordan, Dunn, Lewis & Croft, 2005; Ostelo, de Vet, Knol & van den Brandt, 2004; Stratford, Binkley, Riddle & Guyatt, 1998) included the English version of the 24 point RMQ as at least one of several outcome measures. Lauridsen and colleagues (2006) assessed the MCID of the Danish version of the 23 point RMQ, based on previous cross-cultural validation of this measure (Roland & Morris, 1983). Kovacs and colleagues (2007) evaluated the MCID of the 24 item Spanish version of the RMQ and Grotle and colleagues (2004) used a 24 item Norwegian version. Stratford and colleagues (1998) and McNair et al (2007) were the only studies to investigate the MCID of only the RMQ and LLTQ respectively, with no comparisons made to other outcome measures within the same study. Other outcome measures for which the MCID was investigated within the remaining seven studies included the Oswestry Disability Index, Visual Analogue Scale, Pain Numerical Rating Scale, Quebec Back Pain Disability Questionnaire, SF36, RMQ-18 point version, Physical Functioning Scale, the Main Complaint scale, Low Back Pain Rating Scale and the Waddell Disability Index.

The GPE scales utilised as the anchor to detect minimal change varied widely across the studies. Four of the studies (Beurskens, de Vet & Koke, 1996; Davidson & Keating, 2002; Lauridsen et al, 2006; Ostelo, de Vet, Knol & van den Brandt, 2004) employed a 7 point version, ranging from 'completely recovered', to 'worse' or 'vastly worse'. Other GPE scales employed included a 4 point (Kovacs et al, 2007), 5 point (McNair et al, 2007), 6 point (Jordan, Dunn, Lewis & Croft, 2002), and 15 point GPE scale (Stratford, Binkley, Riddle & Guyatt, 1998). Two of the reviewed studies used additional external criteria to indicate meaningful change (Lauridsen et al, 2006; Stratford, Binkley, Riddle & Guyatt, 1998). Grotle, Brox and Vollestad (2004) was the only study included within the review to compare the MCID based on different external criteria, with the researchers using a 6 point GPE or an expected clinical course as the criterion of change.

There was a wide variation in testing intervals between baseline and follow up to estimate the MCID. The follow up intervals between testing points across the RMQ studies ranged from 1 week (Lauridsen et al, 2006) to 6 months (Jordan, Dunn, Lewis &

Croft, 2006). McNair and colleagues assessed subjects using the LLTQ at baseline and following 1-6 weeks of treatment (McNair et al, 2007).

2.7.3.6 Distribution based statistical methods employed

As the current review has not specifically selected studies pertaining to the statistical responsiveness of the respective outcome measures, conclusions regarding such responsiveness of the RMQ and LLTQ cannot be made. However, it has been widely recommended that studies of the MCID also employ distribution based measures (Crosby, Kolotkin & Williams, 2003; de Vet et al, 2006; Guyatt et al, 2002). Stratford and colleagues (1998) were the only study whereby distribution based methods of analyses from data gathered from scores within the same study were not employed, rather they cited statistics obtained from RMQ score change data from a previous study (Stratford et al, 1996). Various distribution based analytical methods were employed within the reviewed studies to assess statistical change or measurement of error. These methods included calculations of effect size, SRM, SEM and MDC. Six of the reviewed studies (Beurskens, de Vet & Koke, 1996; Davidson & Keating, 2002; Grotle, Brox & Vøllestad, 2004; Lauridsen et al, 2006; McNair et al, 2007; Ostelo, der Vet, Knol & van der Brandt, 2004) included calculation of the SRM or effect size and five studies (Davidson & Keating, 2002; Jordan, Dunn, Lewis & Croft, 2006; Kovacs et al, 2007; McNair et al, 2007; Ostelo, der Vet, Knol & van der Brandt, 2004) included in the review made calculation of the SEM or the MDC of the measures to give an estimate of measurement error.

2.7.3.7 Minimal clinical important difference results

The MCID of the RMQ and LLTQ generated within the studies were presented in various forms. The MCID in four of the studies pertaining to the RMQ (Beurskens, de Vet & Koke, 1996; Davidson and Keating, 2002; Grotle, Brox & Vøllestad, 2004; Ostelo, de Vet, Knol & van der Brandt, 2004) were given as absolute change scores, ranging from 2.5 to 8.6 points, with intention that these values could be applied to any baseline score. Grotle and colleagues (2004), comparing acute and chronic LBP subjects, generated a higher MCID for the acute cohort (2.5 points) compared to the chronic group (1.5 points). McNair et al (2007), in establishing the MCID of the LLTQ, used an additional approach to ROC curve analysis by applying a mean calculation of anchor, distribution and likelihood ratio results to generate an MCID of 4 points for both domains of the outcome measure. The remaining studies presented the MCID as

ranges based on initial scores (Stratford, Binkley, Riddle & Guyatt, 1998; Kovacs et al, 2007) or as a percentage change scores (Jordan, Lewis, Dunn & Croft, 2006; Lauridsen et al, 2006). Stratford et al (1998) developed score ranges for the MCID based on different subsets of the population according to initial RMQ scores. The MCIDs calculated by the researchers varied according to the initial scores. Estimates of MCIDs were 2, 4, 5, 8 and 8 RMQ points, for initial scores of 0-8, 5-12, 9-16, 13-20 and 17-24 points respectively. Kovacs and colleagues (2007) also presented the MCID for the RMQ based on initial scores of subjects, calculating MCIDs of 2.5-6.8 points for initial RMQ scores under 10 and 5.5 to 13.8 points for subjects with initial scores above 15. In contrast, two of the reviewed studies (Jordan, Dunn, Lewis & Croft, 2006; Lauridsen et al, 2006) calculated percentage score changes, designed to take into account any baseline RMQ score, calculating 30% and 38% reductions respectively to represent the MCID.

2.7.4 Discussion

2.7.4.1 Quality assessment

A moderate variation between study methodologies was found. Variance in subjects, test-retest intervals, external criteria employed to define meaningful change, statistical approaches and forms of presenting the MCID were particularly evident. The following sections discuss key areas identified within the appraisal and the influence of these variables on both generating the MCIDs for the two outcome measures and applying them to a wider or specific clinical population.

2.7.4.2 Study participant selection

There was a wide range of subjects used across the reviewed studies including acute LBP, LBP of no defined duration, chronic LBP and post-surgical conditions, potentially limiting the ability to apply the findings to an individual or a specific population. This is particularly important in the assessment of the MCID of an outcome measure as limitation in physical function and levels of change deemed meaningful have been reported to vary across different LBP populations (Grotle, Brox & Vollestad, 2004; Kovacs et al, 2007; Revicki, Hays, Cella & Sloan, 2008). In addition, several authors have identified differences in clinical setting and type and duration of treatment received as potential influences on MCIDs generated for the RMQ (de Vet et al, 2006; de Vet et al, 2007; Demoulin, Ostelo, Knotterus & Smeets, 2010; Stratford & Riddle,

2005). The findings of the current literature review suggest that the influence of subject variation on the MCID of the RMQ and LLTQ has not been widely investigated and are subsequently largely unknown. The findings of two studies (Grotle, Brox & Vollestad, 2004; Kovacs et al, 2007) within the review where the effect of symptom duration on the MCID of the RMQ was investigated, suggest that there is little influence of LBP duration on the MCID, consistent with more recent reports (Demoulin, Ostelo, Knotterus & Smeets, 2010). No other studies included in the current review directly investigated the effect of this variable on the MCID of either outcome measure within the same study, therefore inferences regarding the extent of influence is limited.

2.7.4.3 Use of the Global perceived effect scale

Various definitions of minimal change were used based on a variety of GPE scales employed within the appraised studies. The extent of the influence of this variable has not been elucidated, however standardisation in the use of GPE scales regarding the number of items and phrasing used, and the item on the scale used to define minimal change has been identified as an area of improving outcome measure research (Lauridsen et al, 2007; Terwee et al, 2010). The issue of how to categorise change scores into improved and non-improved groups has also been highlighted as a potential problem for evaluating the MCID of outcome measures (Revicki, Hays, Cella & Sloan, 2008). As all studies included within the review dichotomised subject GPE responses into those at or above the minimal level of improvement and those not indicating improvement, outcome measure change scores from subjects reporting higher levels of improvement than the minimum cut off point are included within the analysis of the MCID. As a result, the MCID established using this method may not truly represent the minimal level of meaningful improvement.

2.7.4.4 Influence of baseline scores

The extent to which baseline outcome measure scores influence the resultant MCID of the RMQ was investigated within four of the reviewed studies (Jordan, Dunn, Lewis & Croft, 2004; Lauridsen et al, 2006; Kovacs et al, 2007; Stratford, Binkley, Riddle & Guyatt, 1998). These findings suggest that subjects with lower levels of physical function at baseline, represented by a larger score, require higher score changes to show clinically important improvement than those with higher physical functioning capacity at baseline.

There are several explanations for the association between baseline scores and the MCID. Firstly, there is anticipated regression to the mean where a variable that is extreme on its first measurement will tend to be closer to the centre of the distribution on a later measurement (Barnett, van der Pols & Dobson, 2005), thereby subjects with very high or very low baseline scores will fall closer to the mean score at follow up (Beaton, Boers & Wells, 2002; Copay et al, 2007). Floor and ceiling effects are also a consideration, where subjects whose scores are particularly low or high may not be able to register a change that is above that of the MCID as they may exceed the span of the scale. This must be addressed particularly when considering the MCID in populations where subjects may have lower levels of limitation in physical function initially, therefore cannot exhibit high levels of improvement on the given measure at follow up (Copay et al, 2007; Demoulin, Ostelo, Knotterus & Smeets, 2010). As the RMQ and LLTQ are non-interval scales, the amount of change according to the scores given on the measure may represent different degrees of change. As a result, each point of the score cannot be considered equal (Farrar et al, 2000; Guyatt et al, 2002). The achievement of some tasks, for example, 'getting out of a lounge chair', requires similar movement patterns and loading requirements to other tasks such as 'getting in and out of a car', therefore the achievement of one task may automatically correspond with the achievement of others. In addition, different functional tasks may carry more weighting in regards to the importance for different individuals, and varied clinical populations likely have differing functional requirements. These additional factors may influence the MCID regardless of baseline scores (Revicki, Hays, Cella & Sloan, 2008). Although the literature appears to support the likelihood of an influence of baseline scores on MCIDs, there is currently no formal consensus within the literature as to the extent or nature of this influence, nor how the MCID should be adjusted according to baseline limitation in function (Beaton, Boers & Wells, 2002; Demoulin, Ostelo, Knottnerus & Smeets, 2010; Terwee et al, 2010; van der Roer et al, 2006).

2.7.4.5 Previous Reviews

The LLTQ has not been included within any previous reviews pertaining to responsiveness or MCID, however a recent review including research assessing the MCID of the RMQ was found within the literature search (Ostelo et al, 2008). This review highlighted the constraints in establishing a single working MCID for the RMQ, and recognised the limitation in pooling results from the current evidence base due to the heterogeneity of subject populations, study methodologies, variation in definitions

of important change and implications of the effects of varying baseline scores (Ostelo et al, 2008). The authors, through evidence review and expert consensus, proposed an MCID of 5 points for the RMQ, or a score reduction of 30% from baseline (Ostelo et al, 2008). The review however had several limitations. The heterogeneity of the studies reviewed limits consensus as to the optimal method of generating the MCID and the value of the MCID itself. In addition, the review included studies that did not generate an MCID for the RMQ using anchor based strategies, recognised as important when ascertaining clinically important change of an outcome measure (Beaton, Boers & Wells, 2002; Guyatt et al, 2002; Revicki et al, 2006; Revicki, Hays, Cella & Sloan, 2008).

2.7.5 Summary of Review

Overall, variation in subjects and methodologies between the reviewed studies in addition to varying statistical methods of analysis does not lend favourably to strongly indicate a set value or score to represent the MCID of the RMQ or LLTQ. The current review indicates RMQ reduction scores of anywhere between 2-3 and 8-9 points when separating subjects according to baseline RMQ scores, however a value of close to 4-5 points, or a 30% reduction in RMQ score from baseline is consistent with the reviewed evidence and has been indicated to represent the MCID within other reviews (Ostelo et al, 2008). The LLTQ has been investigated within a population with lower limb conditions, with MCIDs of 4 points generated for both domains. These values offer some clinical guidance as to values indicative of important change when using the RMQ and LLTQ, however the accuracy may be limited according to individual or population applied to, particularly where baseline scores fall close to either end of the scale (Beaton, Boers & Wells, 2002; Ostelo et al, 2008).

2.7.6 Limitations of the Review

This review has several limitations. A meta-analysis was not performed due to the small number of studies and large variation in study methodology. The review excluded studies not published in English and excluded those studies where calculations of MCIDs were not made based on anchor based approaches. There was a potential for introducing reviewer bias with only one reviewer used to search and complete the review of literature.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter was divided into a number of sections. The first section provides the study design. The second section describes subject recruitment procedures, including selection criteria. The third section provides details of the outcome measures employed and highlights the procedures used to collect the data. The final section pertains to the statistical analyses utilised to synthesise study results.

3.2 Study Design

The study was a quantitative prospective longitudinal evaluation of two outcome measures: the RMQ and LLTQ, focusing on limitation in physical function.

3.3 Subjects

3.3.1 Power & Effect Size

Sample sizes required to identify the MCID of an outcome measure are currently unknown (Terwee et al, 2010). Therefore, based on the reliability testing of the primary dependent variable, and the technique described by Tyron (2001) for comparing responsiveness and minimal important difference scores with power set at 0.8 and alpha level of 0.05, the number of subjects required to observe a moderate effect was 60. A rate of 10% was estimated for potential attrition, thus the aim was to recruit 66 subjects.

3.3.2 Subject Recruitment

In accordance with the requirements of the Northern Regional Ethics Committee, subjects with acute LBP were invited to participate at their initial visit to the Avondale Physiotherapy Clinic. This clinic is a private practice operating within the Auckland suburb of Avondale. The patient base for the clinic covers a wide spectrum of the population with primarily musculoskeletal complaints who have been self referred, or referred from local General Practitioners or Specialists.

Patients were screened to assess eligibility according to selection criteria.

3.3.3 Inclusion Criteria

Subjects were aged over 18 years, presenting with an acute episode of LBP of less than 6 weeks duration. LBP was defined as the presence of symptoms between the thoraco-lumbar junction and the sacrum, with or without referred symptoms into the lower limb/s.

3.3.4 Exclusion Criteria

Patients were excluded if there was a poor understanding of verbal or written English or inability to provide informed consent. Pregnant women and patients with symptoms and/or signs of cauda equina syndrome, progressive paresis, fracture, suspected malignancy, infection, rheumatoid arthritis or other inflammatory diseases were also excluded.

3.3.5 Consent Procedure

Once screened for eligibility, potential subjects were given an information sheet (see Appendix 3) outlining the study format and procedures, contact numbers and an informed consent form (see Appendix 4). Subjects were encouraged to ask questions and were assured that participation in the research would not affect the treatment they received and they were able to withdraw from the research at any stage. Once informed consent was given, subjects were then given the outcome measures to complete prior to their initial intervention with the treating therapist.

All subjects received physical therapy intervention for their injury during the study. Because the assessment of treatment effectiveness was not the purpose of our study, the specifics of the intervention are not relevant.

3.3.6 Personal and Demographic Data

Personal demographic, diagnostic and past history data were collected through the normal physiotherapy examination where notes were recorded and stored electronically as per normal practice requirements. These data included gender, age, occupation, time

since onset of LBP, presence of leg symptoms, past history of LBP, time since last episode of LBP, pain based on a 10 point visual analogue scale (VAS) and the clinical or structural diagnosis recorded by the treating therapist.

3.4 Outcome Measure Procedures

Subjects completed the LLTQ (see Appendix 5) and RMQ (see Appendix 6) on three occasions: at initial visit, at the follow up testing point where a level of 'improved' was indicated using the GPE scale (see Appendix 7), and at discharge from treatment or when 'complete recovery' was indicated. The GPE scale was used at every follow up visit for physiotherapy treatment to gauge overall change. A 7 point likert version of the GPE scale was employed as the anchor to measure clinically meaningful change. This rating scale had the response options of : "completely recovered," "much improved", "improved", "slightly improved", "unchanged", "slightly worse" and "much worse". A 7 point scale was chosen as it was considered to be a middle ground between short scales that may lack clear distinction on levels of change, and long scales which were potentially more difficult for subjects and therapists to interpret and understand.

Subjects completed all questionnaires in random order to minimise any potential bias effect from either outcome measure. This procedure involved the toss of a coin.

All questionnaire completion occurred within the clinical setting. This method was chosen over other potential methods such as mailing questionnaires to ensure scores accurately reflected the condition of the subject at that time and to be certain that outcome measures were completed concurrently for appropriate comparisons to be made. The completion of outcome measures in the clinical setting enabled the treating therapist or practice manager to ensure that all items of the measures had been completed. In addition, a higher loss at follow up was anticipated if completion occurred at home with administration by mail.

Following data collection at initial and follow up visits, the completed questionnaires were forwarded to the principal researcher for analysis and data entered into Microsoft Excel 2007 and Windows SPSS (version 15.0 (SPSS Inc., Illinois, USA) for subsequent analyses. All completed questionnaires were assessed for normality to ensure the appropriate statistical tests were undertaken in the subsequent analyses.

Outcome measure scores across testing points were analysed to assess statistical magnitude of score changes and score changes were analysed according to baseline scores when standardised into percentage of maximum scores.

3.5 Investigation A; Content Validity; LLTQ Importance scores

In the analyses of the LLTQ importance rating data, the four possible responses given for the importance of each task were dichotomised as 'not important' or 'some importance' which included grouping of the three other responses. The frequencies of these categories of responses were expressed as percentages for each item.

3.6 Investigation B; Construct Validity; Correlation analysis between RMQ and LLTQ scores

Baseline RMQ scores were reversed and along with LLTQ scores, were standardised into percentages of maximum scores. Linear regression analysis was undertaken to assess the correlation between standardised scores. It was thought that the LLTQ domains would have at least moderate levels of correlation with the RMQ scores. An Intra-class Correlation was used to assess this association. A 2-way random model examining absolute agreement was utilised.

3.7 Investigation C; Statistical Responsiveness; Calculation of effect size and standardised response mean

Effect size was calculated as the mean change between initial and final scores, divided by the standard deviation of the baseline score (Kazis, Anderson & Meenan, 1989). SRM's were calculated as the mean score change between initial and final testing, divided by the standard deviation of the change score (Liang, Fossel & Larson, 1990). Effect sizes are classified as small (0.2), moderate (0.5) and large (>0.8) (Cohen, 1977).

3.8 Investigation D; Measurement of Error: Calculation of standard error of measurement and Minimal detectable change

The following formula was used to calculate the SEM: standard deviation $* \sqrt{(1-R)}$ (Crosby, Kolotkin & Williams, 2003; Davidson & Keating, 2002; Ferguson, Robinson & Splaine, 2002; Jordan, Dunn, Lewis & Croft, 2006; Ostelo, de Vet, Knol & van den Brandt, 2004), where the standard deviation is recorded as the variance of subject scores at baseline mean change scores and R is the measure of reliability adopted from previously established reliability parameters of the two outcome measures.

For the calculation of the SEM of the RMQ, a test-retest reliability parameter calculation of 0.86 was adopted from a previous study by Stratford and colleagues (1996). This value was selected as it was sourced from a similar sample to the current study. It also represents a mid-point of other estimates of reliability from research pertaining to the RMQ (Kopec et al, 1995; Jacob, Baras, Zeet & Epstein, 2001; Johansson & Lindberg, 1998; Stratford, Binkley & Riddle, 2000; Underwood, Barnett & Vickers, 1999). The reliability parameter for the LLTQ was the ICC reported by McNair et al, 2007; 0.96 (ADL Domain) and 0.98 (Recreational Domain). These ICCs were chosen as the reliability estimates for the LLTQ rather than the Cronbach alpha used by McNair and colleagues (2007). As the ICC is considered a more appropriate statistic for examining test-retest reliability through evaluating correlation based upon estimates from analysis of variance (Domholdt, 2005), it has been recommended as the reliability estimate of choice when calculating the MDC (Kopec & Esdaile, 1995; Resnik & Dobrzykowski, 2003; Rocchi et al, 2005).

The following formula was used to calculate the MDC; 1.645 * square root of 2 * SEM (De Vet et al, 2006). For calculation of the MDC, rather than use a 95% confidence interval, several authors have proposed that the 90% confidence interval is sufficient for application to clinical practice and the use of this lower confidence interval may act to compensate for the often overestimated degree of error encountered with a small sample size and potentially small reliability estimate (Davidson & Keating, 2002; de Vet et al, 2006; Rothstein & Echternach, 1993).

3.9 Investigation E; Minimal clinical important difference of the LLTQ; Receiver operating characteristic curve

The data used for the calculation of the MCID of the LLTQ domains was from only change scores of subjects who reported themselves 'improved' at follow up testing. The LLTQ scores of these subjects were compared against 65 subject control change scores derived from a previous study using the LLTQ where subjects reported to be unchanged using a similar criterion of change (McNair et al, 2007). To characterise the LLTQ changes and the GPE results, ROC curves were derived and sensitivity and specificity values were calculated. The area under the curve was calculated together with a 95% confidence interval. The MCID was chosen based on the cut off point nearest the upper left hand corner of the graph giving the highest combination of sensitivity and specificity, with likelihood ratios calculated for each possible MCID.

In a secondary investigation to assess the effect of the baseline scores on the MCID, subject results were differentiated into three groups based on their initial scores for each of the LLTQ domains of below 20, 20 – 30 and scores over 30. The rationale for choosing these categories was based upon personal communication with Prof Peter McNair who led the development of the LLTQ (McNair et al 2007). Based upon his experience with the questionnaire, it was suggested that values less than 20 represent people with severe problems in function while those scoring between 20 and 30 points have moderate limitations and those scoring above 30 points have minor limitation in function. Analysis was repeated based on the change scores of these three groups for both domains. Using the approaches suggested by Hanley and McNeil (1982), z-scores were calculated to compare the ROC curves of the groups.

A regression analysis was performed between initial LLTQ scores and change scores to a level of 'improved' as a tertiary investigation to further assess the influence of scores at baseline.

CHAPTER 4: RESULTS

4.1 Introduction

This chapter is divided into three main sections. Firstly the participants demographic information is provided followed by the responses to the LLTQ and RMQ scores. The remaining sections provide the results of the analyses for the five key investigations.

4.2 Subjects

Seventy-eight subjects were initially recruited, however nine subjects dropped out of the study and did not complete testing. Their reasons were: declining further physiotherapy treatment (1), declining the completion of testing at follow up (1) and failing to return to follow up appointments (4). Additionally, three subjects who completed testing at baseline did not reach a level of at least 'improved' during the study period. The data set for these nine participants has therefore been excluded from subsequent analyses. The demographics and initial scores for the outcome measures were similar between the nine drop-outs and for the remaining sixty-nine subjects who completed testing. Demographic and symptom information is presented in table 4.1.

Table 4.1: Subject and symptom information collected at baseline

Characteristic	N	%
Gender		
Male	35	51
Female	34	49
Age (years)		
Mean	44 ± 14	
Time since onset of Symptoms (Days)		
Mean	10 ± 10	
Presence of leg Symptoms		
Yes	12	17
No	57	83
Previous History of LBP		
Yes	40	58
No	29	42
Time since previous episode (Months)		
Mean	18 ± 27	
Paid Employment		
Yes	48	70
No	21	30
Pain VAS (1-10)		
Mean	5 ± 2	
N = number of subjects	% = percentage of subjects	

The mean duration between onset of symptoms and commencement in the study (first treatment) was 10 days. Fifty-eight percent of all subjects had a history of previous LBP, with a mean duration of 18 months since their last episode.

Upon evaluating subject change using the GPE scale at each follow up, fifty subjects reported to be at the GPE level of 'improved' and therefore repeated completion of the outcome measures. The remaining subjects completed the outcome measures at follow up testing when reporting a level of 'much improved' or 'completely recovered' as data from these subjects was not captured when only a level of 'improved' was reached.

At the final testing point, where subjects were discharged or indicated a level of 'completely recovered' sixty-four subjects completed testing, with five subjects dropping out of the study after testing at a level of 'improved'.

4.2.1 Subject diagnosis

A total of eight clinical or structural diagnoses were made by the treating therapist. Therapist practice experience ranged from 6 months to 5 years. Two of the four treating therapists held post-graduate qualifications in musculo-skeletal physiotherapy. The two most common diagnoses were 'disc sprain/strain' (n=24) followed by 'lumbar strain/sprain' (n=19). These results are displayed in Table 4.2

Table 4.2: Subject Diagnoses

Diagnosis	N	%
Disc Sprain/Strain	24	35
Lumbar Strain/Sprain	19	28
Lumbar Facet Sprain	6	9
Disc Herniation/Bulge	5	7
Disc Derangement	5	7
Lumbar Instability	4	6
Muscle Strain	2	3
Thoracolumbar Sprain	2	3
Not Reported	2	3

N = number of subjects   % = percentage of subjects

4.3 Time and treatment sessions

The time between the initial visit and the follow up visit where subjects reported a level of 'improved' ranged from 2 to 38 days, with a mean of 8 (SD $\pm$ 6) days. The time between initial testing and 'completely recovery' or discharge, ranged from 3 to 44 days, with a mean of 19 (SD $\pm$ 12) days. The number of physiotherapy treatments received from baseline to 'improved' ranged from 1 to 4 sessions, with a mean of 2 (SD $\pm$ 1). The number of treatments to final testing indicating complete recovery or discharge ranged from 1 to 12 sessions, with a mean of 4 (SD $\pm$ 3).

4.4 Outcome measure responses

4.4.1 Global perceived effect score responses

At follow up, fifty out of the sixty-nine subjects reported that they were 'improved', and of the 64 subjects tested at discharge, subjects reported being 'improved' (4) 'much improved' (23) or 'completely recovered' (37). Four subjects reported worsening of symptoms at follow up visits, however continued treatment and were improved at subsequent visits. One subject reported worsening in symptoms following an improvement, however improved at subsequent visits.

4.4.2 LLTQ Scores over testing points

The mean LLTQ scores at baseline testing were 25 (SD ± 7) and 17 (SD ± 9) for the ADL and Recreational domains respectively. At follow up testing where a level of 'improved' was indicated by fifty subjects, there was a significant increase in mean LLTQ scores by 7 (SD ± 4) and 6 (SD ± 4) points for the respective domains to a mean of 31 (SD ± 6) and 23 (SD ± 9) points respectively ($p<0.05$). Between baseline testing and final testing, there was also a significant increase in mean LLTQ scores by 12 (SD ± 6) points in the ADL domain, and a mean change of 15 (SD ± 8) points in the Recreational domain to 37 (SD ± 4) and 33 (SD ± 6) points respectively ($p<0.05$).

With respect to floor and ceiling effects of the measure, one subject at baseline reported the maximum score in the ADL domain, and 20% of subjects reported the maximum score in the ADL domain at 'improved'. At final testing, 44% and 14% of subjects reported the maximum score of 40 points in the ADL and Recreational domains respectively. No subjects reported the minimum score in either domain at any testing interval.

The following graph displays the means and standard deviations of the LLTQ scores over the three testing points (Figure 4.1).
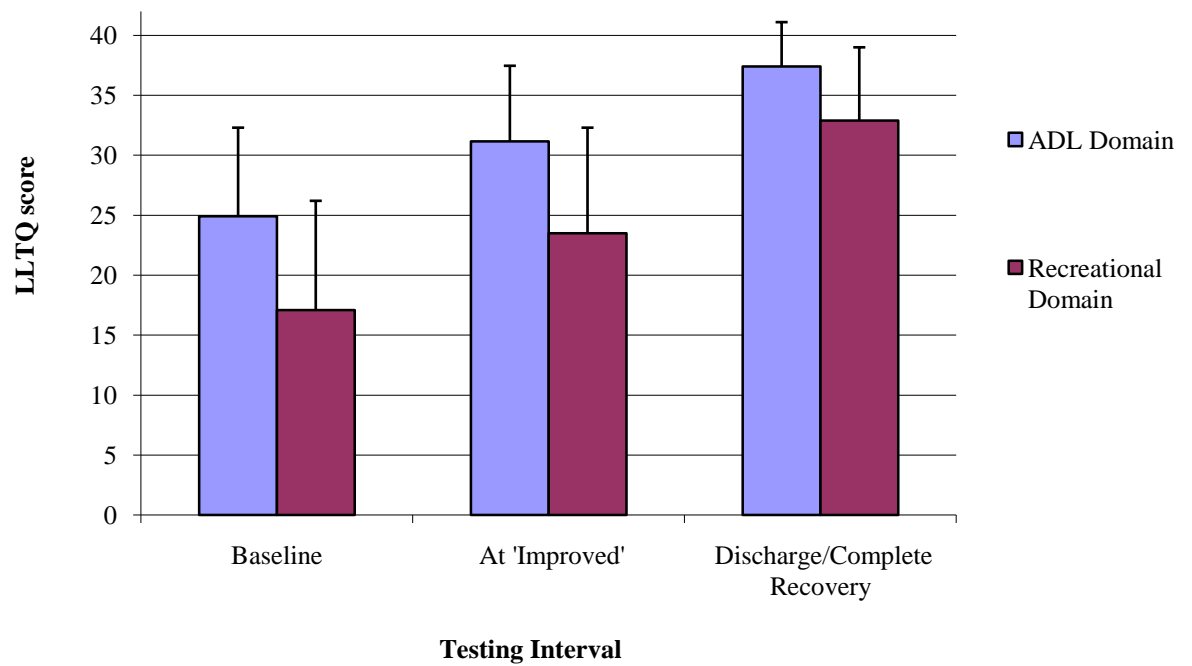
Figure 4.1: LLTQ Mean Scores over Testing Points

4.4.3 RMQ Scores over testing points

Scores for the RMQ reduced significantly from initial testing by a mean of 4 (SD ± 2) points to a mean of 7 (SD ± 4) where a level of 'improved' was met on the GPE (p<0.05), and mean change of 9 (SD ± 5) points between initial testing to a mean of 2 (SD ± 3) points at complete recovery or discharge (p<0.05). No subjects recorded a maximum or minimum score on the measure at baseline, however twelve percent of subjects recorded the best possible score of 0 points at 'improved' and thirty percent of subjects recorded this score at final testing. These findings over the testing points are displayed in Figure 4.2.
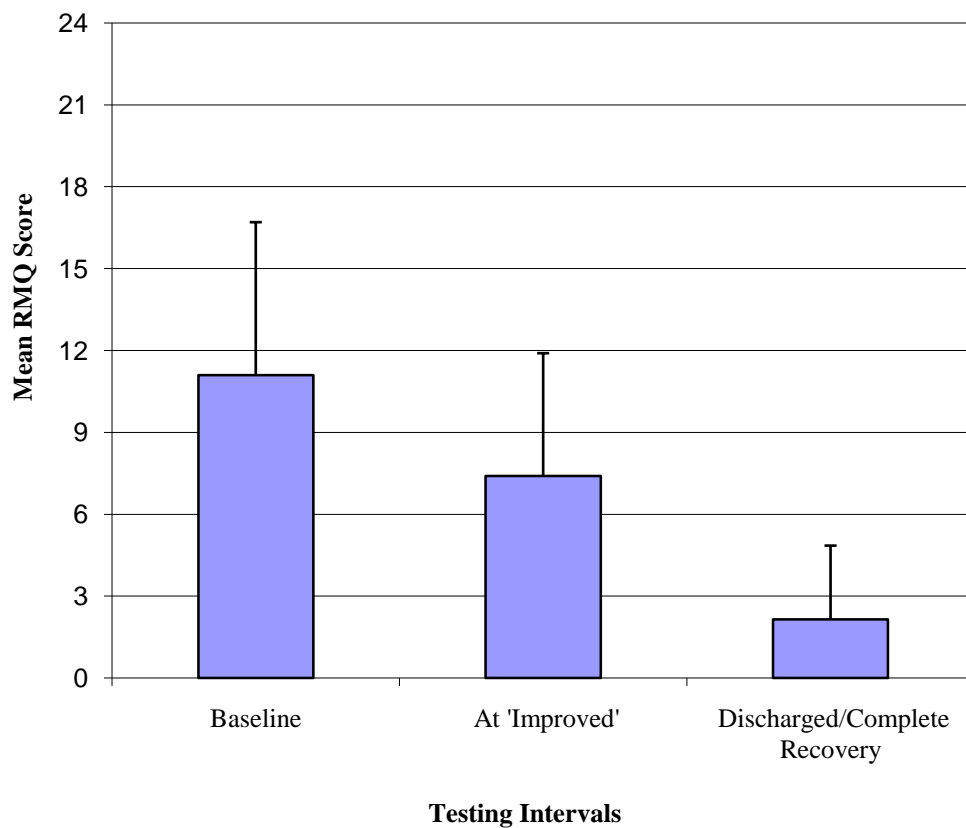
Figure 4.2: RMQ Mean scores over testing points

4.4.4 Outcome measure score changes over testing according to baseline percentage scores

In order to compare outcome measure scores between testing points, RMQ scores were reversed and scores of both measures were standardised into percentages of maximum score. To assess whether subject scores at baseline had an influence on change scores at follow up testing, subjects initial scores were categorised according to percentage quartiles of baseline scores. The following figures display the mean score changes as percentages of maximum scores between baseline and follow up testing for subsets of the population based on their initial scores.
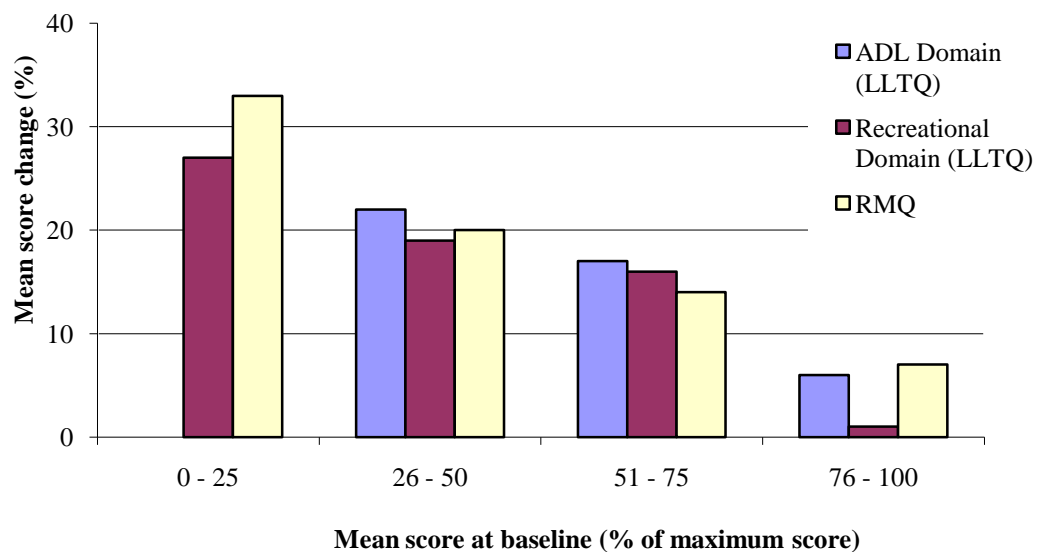
Figure 4.3 Mean percentage change in scores to a level of 'improved' based on initial scores
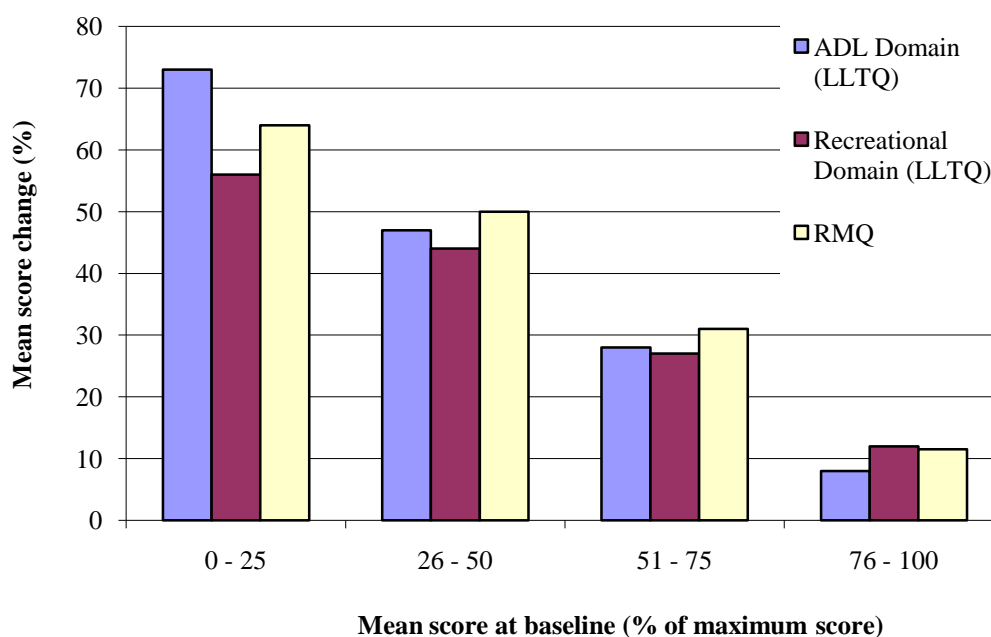
Figure 4.4 Mean percentage change in scores to discharge or 'complete recovery' based on initial scores

Lower LLTQ and higher RMQ scores at baseline, represented by the lower two quartiles of initial scores, demonstrated higher mean change in scores from baseline testing to both follow up points of 'improved' and discharge/complete recovery.

Although demonstrating a higher mean score change over testing, lower physical functioning at baseline also corresponded with lower physical functioning at both follow up testing points. This was particularly apparent with the LLTQ Recreational scores where those with the lowest scores at baseline had the highest mean increases of scores of 27% from baseline to improved, and 56% from baseline to final testing, yet at both 'improved' and 'discharge/complete recovery' those with lower initial scores had overall lower scores than subjects with higher baseline scores. This was apparent where the mean score at discharge was only 71% of the maximum score, comparable to that of 98% for the group with the highest initial scores.

The results of the linear regression to further assess the relationship between initial scores and score changes are presented in Investigation E (section 4.9).

4.5 Investigation A; Content Validity; LLTQ Importance Scores

Subjects responses to the importance of tasks in the two domains of the LLTQ are presented as percentages of subjects who regarded each task to be of at least some importance. These responses are presented in Table 4.3.

Table 4.3: Percentage of Subjects Regarding LLTQ Tasks as of some Importance

LLTQ Item	Subjects rating task as of at least some importance (%)
ADL Domain	
Walk for 10 minutes	99
Walk up and down 10 steps (1 flight)	97
Stand for 10 minutes	90
Stand for a typical work day	87
Get on and off a bus	75
Get up from a lounge chair	90
Push or pull a heavy trolley	86
Get in and out of car	94
Get out of bed in the morning	96
Walk across a slope	78
Recreational Domain	
Jog of 10 minutes	65
Pivot or twist quickly while walking	81
Jump for distance	48
Run fast/Sprint	52
Stop and start moving quickly	84
Jump upwards and land	74
Kick a ball hard	46
Pivot or twist quickly while running	61
Kneel on both knees for 5 minutes	75
Squat to the ground/floor	83

At least three-quarters of all subjects regarded all tasks in the ADL domain to be of at least some importance. The two tasks in which the least number of subjects responded to be of at least some importance in the ADL domain was getting on and off a bus (75%) and walking across a slope (78%). The tasks regarded by most subjects as being of at least some importance was walking for ten minutes (99%) and walking up and down ten steps (97%). In the Recreational domain, the two tasks in which the least number of subjects regarded to be of at least some importance was the tasks of kicking a ball hard (46%) and jumping for distance (48%). The two tasks in which the highest number of subjects reported as being of at least some importance in this domain were stopping and starting moving quickly (84%) and pivoting or twisting quickly while walking (81%).

4.6 Investigation B; Construct Validity; Correlation analysis between RMQ and LLTQ scores

The following two figures display the results of the linear correlation between the reversed RMQ scores and scores of the LLTQ domains at baseline when standardised as percentages.
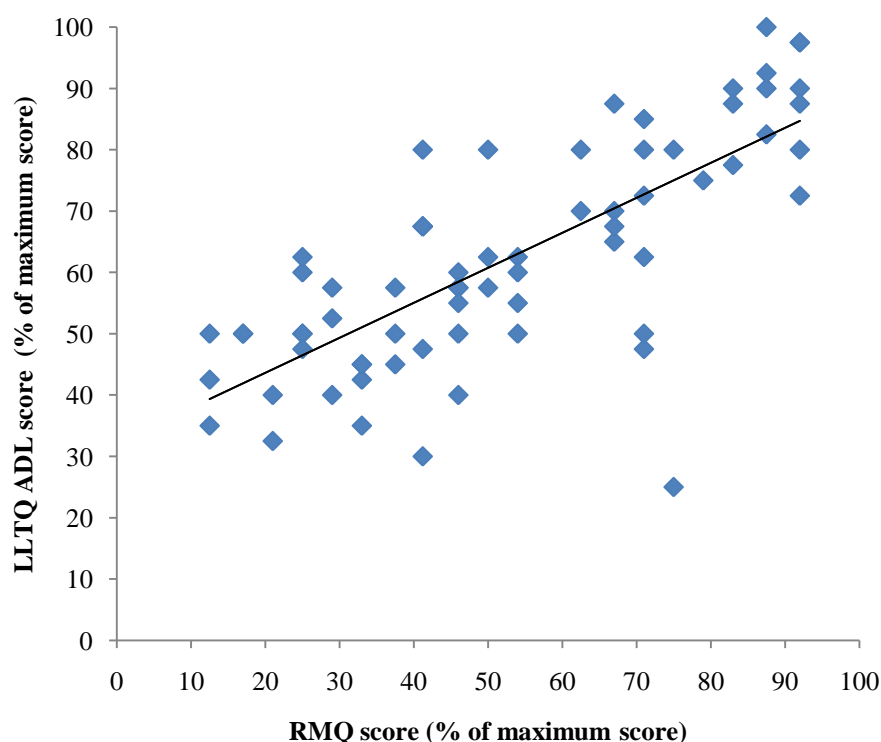
Figure 4.5: Correlation of RMQ and LLTQ ADL domain scores at baseline
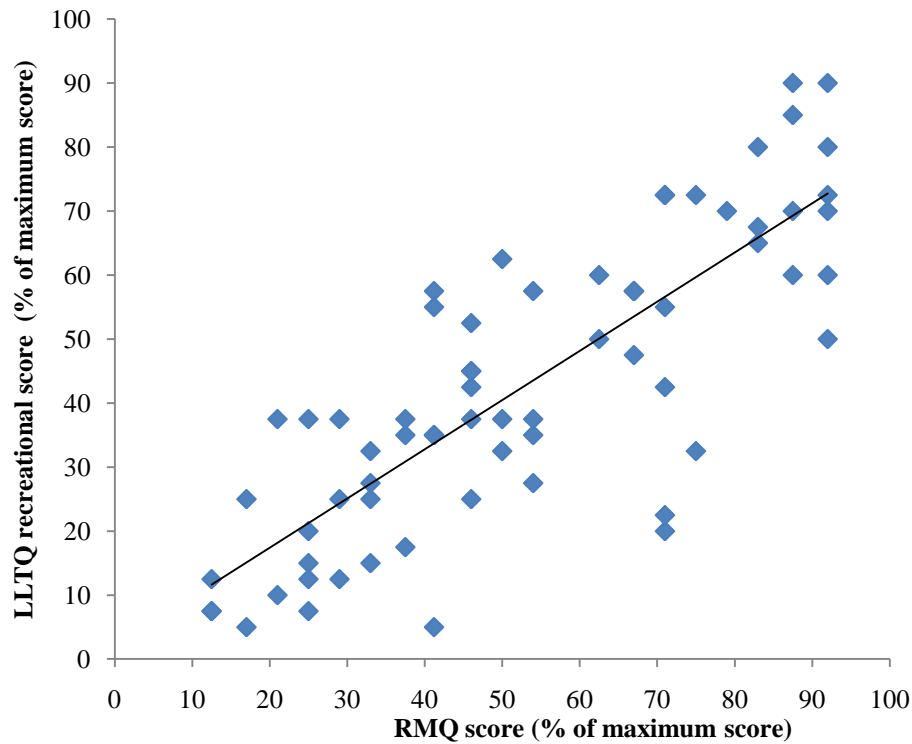
Figure 4.6: Correlation of RMQ and LLTQ Recreational domain at Baseline

The slope of the line of best fit for correlation between the RMQ and LLTQ ADL baseline scores was 0.57 ± 0.06, and the RMQ baseline scores corresponding with the LLTQ Recreational domain generated a slope of 0.77 ± 0.67. The y intercept of the slope of best fit for the ADL domain of the LLTQ was 32.2% on the RMQ scale, and 2% of the RMQ for the Recreational domain. The difference between the line of best fit for both domains of the LLTQ when correlated with the RMQ was significant ($p < 0.03$). An r squared value of 0.56 was generated for the RMQ correlation with the ADL domain of the LLTQ and 0.67 between the RMQ and Recreational domain.

The ICCs and their lower confidence intervals between the RMQ scale the LLTQ were 0.66 (0.39) and 0.75 (0.41) for the ADL and Recreational domains respectively.

4.7 Investigation C; Statistical Responsiveness: Calculation of effect size and standardised response mean of change scores

The effect sizes of the LLTQ ADL and Recreational domains were 1.6 and 1.7 respectively and the effect size of the RMQ was 1.6. The SRM's were 1.5 for both domains of the LLTQ and 1.8 for the RMQ (Table 4.4).

4.8 Investigation D; Measurement of Error: Standard error of measurement and Minimal detectable change

The raw and percentage SEMs and MDCs of the LLTQ and RMQ scores are presented in table 4.4. The SEM generated for the LLTQ were 1.1 and 0.9 for the ADL and Recreational domains respectively, and MDC's were 2.5 and 2.1 respectively. The RMQ scores generated an SEM of 1.4 and MDC of 3.3 points (Table 4.7).

Table 4.4: Results of Distribution Based Statistical Analyses

Outcome Measure	SEM	%SEM	MDC	%MDC	Effect Size	SRM
LLTQ (ADL)	1.1	2.7	2.5	6	1.6	1.5
LLTQ (Recreational)	0.9	2.2	2.1	5	1.7	1.5
RMQ	1.4	5.8	3.3	15	1.6	1.8

SEM% and MDC% = percentage of maximum outcome measure score

4.9 Investigation E; Minimal clinical important difference of the LLTQ; Receiver operating characteristic curve

Area under the ROC curves were 0.95 ± 0.02 with a 95% confidence interval of $0.91 - 0.99$ and 0.95 ± 0.02 with a confidence interval of $0.9 - 0.99$ for the ADL and Recreational domains respectively (Figure 4.7).
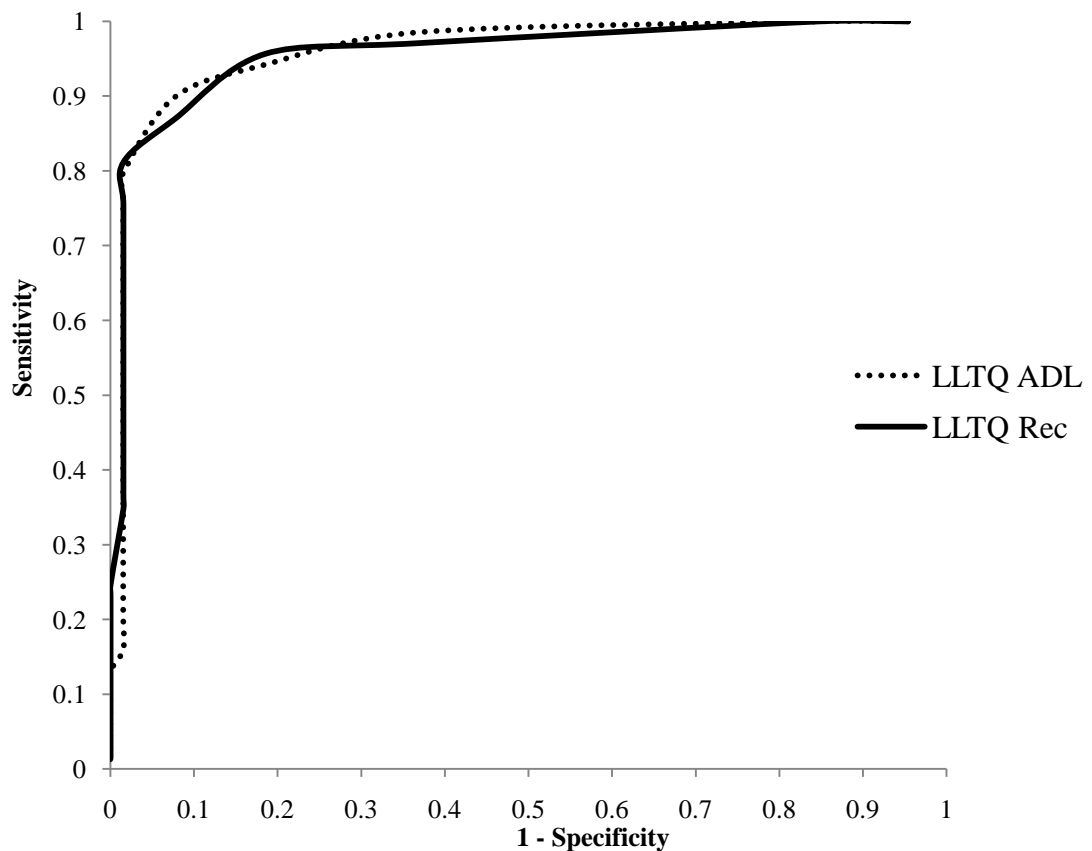
Figure 4.7: The ROC Curves for LLTQ

For ADL change scores between 1.5 and 4 points, sensitivity ranged from 92% to 74%, and specificity from 82% to 98%. For cut off points between 1.5 and 4.5 in the Recreational domain, sensitivity and specificity ranged from 96% to 70% and 82% to 98% respectively. For the ADL domain, the likelihood ratios were above 10 for a cut off point above 2.5, with the maximum likelihood ratio of 48 for a cut off point above 4 points. For the Recreational domain, the likelihood ratios were also above 10 for a cut

off point of 2.5 points, with the maximum likelihood ratio of 49 for an observed cut off score greater than 3.5 points.

Analysis repeated for groups based on initial scores demonstrated areas under the ROC curve of at least 0.97 and a cut off of 2.5 points for subjects with scores below 20 and 20 − 30 points in both domains with likelihood ratios above 10. There was no significant differences between ROC curve results for ADL and Recreational domain initial scores of less than 20 and 20-30 (ADL z-score = 0.71; Recreational z-score = 0.37). There were insufficient subject numbers with initial scores over 30 points to make comparison with the other data categories.

An r squared value of 0.25 was generated for the association between lower LLTQ ADL domain initial scores and higher change scores to a level of 'improved'. The r squared value between the initial scores and change scores for the Recreational domain also demonstrated a negative relationship between initial score and change score with an r squared value of 0.18. A Grubbs test for the Recreational domain regression showed that there was a significant outlier with a score of 32. As this subject had a very low initial score, their change score could lever the line of best fit and hence the analysis was repeated without this data point. The findings showed that the r squared value was reduced from 0.18 to 0.14.

CHAPTER 5: DISCUSSION

5.1 Introduction

This chapter is divided into three main sections. The first section discusses the results pertaining to the subject demographics and outcome measure results. The remaining sections discuss the results of the findings for the five study purposes, followed by a discussion of the limitations of the current study.

5.2 Subjects

The gender ratio, mean age and work status of the subjects were comparable to other baseline demographic data for acute LBP populations (Carey et al, 1995; Dettori et al, 1995; Fritz, Delitto & Erhard, 2003; Rozenberg et al, 2002; Sieben, Valeyen, Tuerlinckx & Portegijs, 2002). The proportion of subjects with a past history of LBP was fifty-eight percent. Other studies using acute LBP populations report rates of fifty to sixty-four percent (Fritz, Delitto & Erhard, 2003; Grotle et al, 2007; Rozenberg et al, 2002;), indicating the recurrent nature of the condition. In the current study, seventeen percent of subjects presented with leg pain in addition to their LBP complaint, which was similar to rates reported in other studies (Fritz, Delitto & Erhard, 2003; Stratford, Binkley, Riddle & Guyatt 1998).

The mean duration from onset of symptoms to the initial testing of 10 days was comparable (8.1 to 12 days) to other studies (Grotle, Brox & Vollestad, 2004; Grotle et al, 2007; Hides et al, 2001; Kovacs et al, 2007; Stratford, Binkley, Riddle & Guyatt, 1998). The mean VAS pain score was also consistent with other reports of baseline pain levels in acute LBP populations (Carey et al, 1995; Grotle et al, 2007; Rozenberg et al, 2002).

5.3 Outcome measure results

5.3.1 LLTQ and RMQ initial scores

The mean RMQ score of 11 (SD ± 5.6) of subjects at initial scoring was consistent with baseline measures from other studies using an acute LBP population, where mean scores have varied from 9 to 13 (Dettori et al, 1995; Carey et al, 1995; Hides, Richardson & Jull, 1996; Sieben, Valeyen, Tuerlinckx & Portegijs, 2002; Grotle, Brox & Vollestad, 2004; Grotle, Brox & Vollestad, 2005). Employing an acute population, it was anticipated that RMQ scores at baseline would be higher than in a chronic subject group as limitation in physical function has been shown to rapidly improve in those with acute LBP within the first 4-6 weeks (Pengel, Herbert, Maher & Refshauge, 2003; Kovacs et al, 2004). This is supported by studies using the RMQ measure in chronic LBP populations with mean RMQ scores of 4 (Rydeard, Leger & Smith, 2006) and 7 (Hides, Gilmore, Stanton & Bohlscheid, 2008).

It was anticipated that as the current study included subjects with a symptom duration of up to 6 weeks, many subjects may have already experienced significant levels of improvement in function prior to seeking treatment. Despite this, baseline RMQ and pain measures did not differ significantly from previous research using more acute populations with a shorter duration of symptoms (Dettori et al, 1995; Grotle et al, 2007; Hides, Richardson & Jull, 1996; Sieben et al, 2002).

Higher levels of limitation in the Recreational domain were anticipated due to the higher loading demands of the tasks (McNair et al, 2007). When comparing the LLTQ ADL and Recreational domain scores at baseline, highest levels of limitation of physical functioning were exhibited in the Recreational domain with a mean score of 43% of maximum score, with lower levels of limitation demonstrated for the ADL domain with a mean of 63% of maximum score. As this is the first study to investigate the LLTQ in an acute LBP population, comparisons with other data from previously published studies using a similar population are not possible. However, in a population with lower limb conditions, McNair and colleagues (2007), reported similar differences at baseline between the two domains of the scores, where higher levels of limitation in physical function were reported for the Recreational domain (McNair et al, 2007). This may

indicate similarity between the limitations of physical function experienced by subjects with acute LBP and those with lower limb conditions. When contrasting these mean percentage LLTQ scores at baseline to those of the RMQ, the mean RMQ score (53%) was half-way between the baseline mean scores for the two domains of the LLTQ. The observed difference of 10% between the RMQ score and ADL score at baseline may be attributable to the nature of the scoring systems rather than differences between functional tasks assessed within the items of the scales. The dichotomous nature of scoring for the RMQ is unable to take into account the partial achievement of tasks as is possible using the LLTQ, therefore subjects may have indicated the inability to achieve items on the RMQ scale which they were encountering limitation with, but marked similar tasks on the ADL domain of the LLTQ as partially achieved, and thus had higher percentage scores at baseline.

The baseline scores for the LLTQ reveal potentially advantageous implications for practical use in LBP populations. With the current subject population scoring the domains of the measure as both higher and lower than the RMQ in regards to physical function at baseline, this may indicate the ability of the two domains of the LLTQ to more accurately measure limitation in physical function in subject populations with widely varying baseline physical functioning capacity. This also demonstrates the potential in using the outcome measure across varying clinical populations who may have higher or lower physical functional requirements such as athletes, or the elderly where requirements of physical functioning differ. For example, where limitation in physical function associated with higher loading recreational based activities is anticipated, such as an athletic population, the Recreational domain may provide a more practical measure to identify the limitation of such individuals and groups. The converse may also be appropriate concerning individuals and groups with higher limitation in physical functioning where the achievement of basic tasks of daily living are affected. In these cases, the ADL domain of the LLTQ may offer a more practical and sensitive measure of functioning.

5.3.2 LLTQ and RMQ change scores at follow up

Outcome measures were assessed at baseline and again where subjects reached a level of 'improved' on the GPE scale employed as the anchor. Follow up visits for treatment were at the discretion of the treating therapist, therefore the actual time between initial testing to where a level of 'improved' was indicated may have been significantly less than the reported mean of 8.3 days. In addition, the treating therapist may have used shorter follow up periods between treatment sessions for patients with lower levels of physical function than for those with less limitation, potentially influencing results. Despite this, the current study has demonstrated significant score changes in both outcome measures between initial scoring and where subjects reported to be 'improved' on the GPE scale ($p<0.05$). This concurred with past research where acute LBP populations have exhibited rapid improvement in physical function as demonstrated with RMQ scores regardless of treatment protocol (Kovacs et al, 2004). This may also explain the relatively short time frame of 8.3 mean days between baseline and where subjects had reached a level of 'improved'.

The change scores in the current study for the LLTQ of 7 (SD ± 4) and 6 (SD ± 4) points for the ADL and Recreational domains respectively are lower than those reported for populations with lower limb injuries (McNair et al, 2007). McNair and colleagues reported a mean increase in scores of 10 (SD ± 8) and 8 (SD ± 7) for subjects who were 'improved' for the respective domains using a re-test period of 7 to 10 days (McNair et al, 2007). This discrepancy in reported change however should not be interpreted as the measure being less responsive or valid within an acute LBP population. The subject population within the study by McNair and colleagues (2007) reported higher levels of limitation of physical function at baseline, therefore higher levels of change would be anticipated at follow up due to a regression toward the mean (Vickers & Altman, 2001). In addition, the authors included post-surgical subjects where rapid improvement was anticipated within several days and the authors adopted a specific retest period whereby they used change scores from all subjects who reported being at or above the level of minimal change. This differed from the current study which retested subjects at varying time periods according to when they reported a level of 'improvement'.

The RMQ scores from baseline to where subjects met a level of 'improved' in the current study had a mean score change of 4 (SD ± 2) points. As this was the first study

to investigate RMQ scores between baseline and a GPE level of only 'improved', it was anticipated that change scores would be lower than if using subjects reporting levels of 'improved' or better. This was confirmed when comparing findings to the study by Stratford and colleagues (1998) who recorded a mean score change of 7.6 RMQ points for subjects reporting any change above a level of 'improved' on a GPE scale following 3-6 weeks of treatment, and other studies where RMQ score changes between baseline and any level above 'improved' have ranged from 5.7 to 8.1 RMQ points (Beurskens, de Vet & Koke, 1996; Grotle, Brox & Vollestad, 2004; Jordan, Dunn, Lewis & Croft, 2006; Kovacs et al, 2007). Despite smaller change scores when using only a level of 'improved' at retesting within the current study, score changes were significant (p<0.05). The importance of re-testing subject scores at the point of 'improved' on the GPE scale was so that analysis for the primary purpose of the study could be made; calculating the MCID of the LLTQ using follow up scores from subjects reporting only a level of 'improved' on the GPE scale as this was defined as the minimal level of change.

5.4 Score changes according to initial scores

Mean change in both LLTQ domains and RMQ scores across testing points varied according to baseline scores. Higher limitation in physical function at baseline, as indicated by a higher RMQ score or lower LLTQ score, was associated with a higher mean change between initial scoring and both testing points, demonstrated through higher percentage change scores and significant correlation being demonstrated between lower levels of function at baseline and higher change scores. This is consistent with literature pertaining to outcome measures of limitation in physical function which report higher change in scores between testing associated with a higher limitation at baseline (Copay et al, 2007; Crosby, Kolotkin & Williams, 2003; van der Roer et al, 2006). There are several potential reasons for this finding. Firstly, with the use of an acute LBP population, it was expected that subjects with high levels of limitation would improve significantly and rapidly as has been demonstrated in similar subject populations (Carey et al, 1995; Coste et al, 1994; Koes, van Tulder & Thomas, 2006; Pengel, Herbert, Maher & Refshauge, 2003). Secondly, there is an anticipated regression toward the mean in follow up testing (Vickers & Altman, 2001). In a non-clinical population, it would be expected that individuals would score at or near the maximum LLTQ score of 40, and 0 points on the RMQ to indicate no limitation in physical function. Therefore at

follow up where subjects reported improvement, it was anticipated that scores further from these 'normal' scores would exhibit a greater change due to a regression toward normal values (Beaton, Boers & Wells, 2002; Copay et al, 2007; Vickers & Altman, 2001). Lastly, the lower levels of mean change exhibited by those with initially higher LLTQ scores or lower RMQ scores may be due to a ceiling and floor effect of the respective measures. Subjects demonstrating the highest levels of physical functioning at baseline, as given by scores within ten points of the maximum LLTQ scores, or within six points of the lowest RMQ score, demonstrated the lowest mean change in scores over testing points as they were already close to the maximum possible scores, therefore had less capacity to demonstrate change on the respective measures than those with higher limitation at baseline.

Although mean percentage change scores differed according to initial scores and a significant negative relationship was demonstrated between these initial scores and change scores, regression analysis for the LLTQ domains indicated that this relationship was not strong, and r squared values to explain the amount of variance in this relationship suggested that initial scores of the LLTQ do not play a large role in the variance in change scores (18-25%). There are several potential reasons for this. The relatively small sample size and few initial scores over 30 points for both domains may have limited the statistical association that was anticipated, and the large accumulation of scores within the middle scoring range may have skewed results. As the domains were assessed separately, there was less of a spread of scores as the separate domains evaluate a set of tasks which are relatively uniform in relation to the level of loading or difficulty required. If domains were added to give an overall score, a stronger correlation between initial scores and change scores may have been generated from a larger spread of scores (Cornell & Berger, 1986).

Subjects with higher levels of limitation in function at baseline also demonstrated higher levels of limitation in function at follow up and discharge. This was most notable in the Recreational domain of the LLTQ. The nineteen subjects scoring initially less than 10 out of 40 points in this domain at baseline demonstrated a mean score of 28 out of 40 at discharge. This mean score is considerably lower than the Recreational domain mean score for all subjects at discharge of 33 out of 40 points. These findings demonstrate ongoing limitation in physical function at discharge for those who are more limited at baseline. This may be due to lower expectations of overall functional recovery

on the part of the patient or clinician for those who are initially worse off, whereby higher limitation is demonstrated when they consider themselves improved and upon discharge from treatment (Myers et al, 2008). This may have important practical implications for the use of the LLTQ in LBP populations within clinical settings and research. Limitation in physical function detected using the Recreational domain of the LLTQ at discharge highlights the strengths of this outcome measure to accurately reflect ongoing problems those with LBP commonly experience (Hides, Jull & Richardson, 2001). This may assist the clinician in identifying areas of persistent limitation in physical function to aid the development of appropriate intervention strategies particularly towards later stages of the injury or in discharge planning. Furthermore, identifying ongoing functional problems prior to discharge may have the potential to reduce the risk of progression to long term limitation in physical function or symptom chronicity, which is widely considered a major issue in the clinical management of LBP (Ehrlich, 2003).

5.5 Investigation A; Content Validity; LLTQ Importance Scores

The responses to the importance category of the LLTQ highlight the functional priorities of the study population and also highlight differences in content between the two domains. All ten tasks in the ADL domain were considered to be at least of some importance to at least seventy-five percent of subjects, demonstrating the priority of such tasks for individuals with acute LBP. These findings are consistent with other research regarding important functional tasks for LBP populations based on the activities and participation component of the ICF model (Cieza et al, 2004; WHO, 2001). Linking of the items of the LLTQ with that of the ICF core sets for LBP model (Stier-Jarmer, Cieza, Borchers & Stucki, 2009) within section 2.5.5 has also demonstrated that the LLTQ includes more than fifty percent of the tasks deemed important in LBP populations. This contrast to the ICF model also highlights that the LLTQ covers function relating to the use of transport and tasks relating to major life areas of work, employment and community and civic life, important daily living functions according to this model which are not included within the RMQ (Pengel, Refshauge & Maher, 2004; Sigl et al, 2006). Regarding the Recreational domain of the LLTQ, more than half of all subjects reported nine out of the ten tasks as being of at least some importance. This finding is congruent with other studies which have demonstrated that social and sporting recreational based activities are important areas of

function for those with LBP (Cieza et al, 2004; Stier-Jarmer, Cieza, Borchers & Stucki, 2009). Such tasks, addressed within this domain, are important areas of function that are also excluded from currently employed outcome measures for LBP populations, including the RMQ (Pengel, Refshauge & Maher, 2004; Sigl et al, 2006).

The findings of this investigation contributes to the evidence suggesting that both activities of daily living and recreational based functional tasks are an integral aspect of assessment of limitations of physical function in LBP populations. The results also indicate that the LLTQ has strong content validity when contrasted to the items outlined within the ICF activities and participation model which have been identified as important in the assessment of those with LBP (Stier-Jarmer, Cieza, Borchers & Stucki, 2009). The LLTQ may also offer additional advantage whereby it includes important items relating to major life areas, not met by other commonly employed outcome measures.

5.6 Investigation B; Construct Validity; Correlation analysis between LLTQ and RMQ scores

Correlation and regression analysis of the outcome measure data was undertaken to assess the linear relationship between the RMQ and LLTQ domains at baseline. There is little agreement regarding the level of correlation between measures to indicate clinical relevance, therefore it has been recommended that values generated are used to describe the relationship and agreement of outcome measures rather than attempt to meet an acceptable standard (Bland & Altman, 1999). A stronger correlation was demonstrated between the scores of the RMQ and the Recreational domain of the LLTQ than for the RMQ and ADL domain, with a significant difference found between the lines of best fit ($p<0.03$). In addition, using the r squared statistic, the RMQ percentage scores at baseline demonstrated better prediction of scores of the Recreational domain (67%) than for the ADL domain (57%). The significant difference demonstrated in the relationship of the domains of the LLTQ with the RMQ may have important practical implications as it highlights the differences in the underlying constructs of physical function being assessed by the separate domains. This is consistent with previous findings of correlation between the two domains of the LLTQ directly, by McNair et al (2007) who reported that although the two domains correlated, (0.78), they only shared approximately 61% common variance using a population with lower limb conditions.

Although the LLTQ and RMQ appear to be measuring similar information pertaining to physical function, there a several explanations for only a moderate correlation being demonstrated between the scores. Differences in the functional items assessed, particular wording of items, response options and the use of a dichotomous scale in the RMQ as opposed to a likert scale in the LLTQ, whereby subjects are able to quantify their ability to achieve tasks using a four point scale, may account for the lack of strong co-efficient estimate being generated.

The ICC's calculated within the current study provide a measure of relevant variance between the contrasted outcome measures to give an indication of consistency between the scores (McGraw & Wong, 1996). The advantage of using the ICC is that it gives a composite value of the within and between score variability to represent the consistency of scores across the outcome measures (Bland & Altman, 1990). Although not formally compared, the higher ICC of 0.75 generated for the RMQ score correlation with the Recreational domain of the LLTQ indicated a higher level of consistency and resemblance of scores than with the ADL domain. The small lower bound scores of 0.39 and 0.41 using a 95% level of agreement for the ADL and Recreational domains respectively however may have been influenced by the sample size employed, presence of outliers and variation of scores within the measures (Bland & Altman, 1990).

There are important implications from the findings of this investigation. As construct validity of physical function outcome measures are formed on the comparison with other measures concerning the same construct (Grotle, Brox & Volestad, 2005), the moderate correlation between RMQ scores and the domains of the LLTQ offer support for the construct validity of the LLTQ in a LBP setting. In addition, the observed variation in correlation between the two domains of the LLTQ with the RMQ may highlight the differences in construct between physical functioning involved in activities of daily life and recreational based activities, both assessed through the LLTQ yet both considered important in the overall physical functioning of an individual with LBP (Cieza et al, 2004; WHO, 2001).

5.7 Investigation C; Statistical Responsiveness: Effect size and Standardised response mean

The effect size and SRM were calculated to assess the magnitude of score changes across testing points, as both statistics can be compared directly across outcome measures used within the same sample and study setting, with values being interpreted as large (>0.8), moderate (0.5-0.8) and small (0.2-0.5) (Cohen, 1977, Horner & Larmer, 2006).

Comparisons with past research investigating the effect size and SRM of the outcome measures may be somewhat arbitrary, as values of these statistics will differ according to variation in study and sample characteristics. This is evident when comparing the results of the effect size and SRM of the RMQ generated in the current study of 1.6 and 1.8 respectively to that of previously published reports. Although most studies pertaining to the effect size and SRM of the RMQ report scores over 0.8, they vary widely within the current evidence base from effect sizes of 0.72 to 2.02 (Beurskens, de Vet & Koke, 1996; Kuijer et al, 2004; Frost, Lamb & Stewart-Brown, 2008) and SRMs of 0.55 to 2.02 (Davidson et al, 2002; Grotle, Brox & Vøllestad, 2004; Hare-Mortensen, Lauridsen & Grunnet-Nilsson, 2006; Lauridsen et al, 2002; Ostelo, de Vet, Knol & van der Brandt, 2004; Kuijer et al, 2004; Turner et al, 2003; Underwood, Barnett & Vickers, 1999).

Both domains of the LLTQ demonstrated high effect sizes of 1.6 and 1.7 and SRMs of 1.7 and 2 for the ADL and Recreational domains respectively. McNair et al, (2007) reported very similar results in a population with lower limb conditions, generating an effect size of 1.5 and 1.6 and SRM of 1.4 and 1.3 for the ADL and Recreational domains respectively. The high effect sizes and SRM's indicate a high level of statistical responsiveness of the LLTQ when applied to both populations with acute LBP and those with lower limb conditions. Additionally, the effect size and SRM calculations were similar for both domains of the LLTQ, which also indicates a high level of statistical responsiveness when using the domains separately or together as a measurement of limitation of physical function.

5.8 Investigation D; Measurement of Error: Standard error of measurement and Minimal detectable change

In order to estimate the reliability of change of the outcome measures and to establish the minimal amount of change in the score that reflects true change above levels of measurement error, the SEM and MDC of the outcome measures were generated.

As the SEM and MDC is displayed in the same units as the outcome measure, direct contrast of this statistic between the RMQ and LLTQ is limited, therefore percentage SEMs and MDCs were calculated to facilitate comparison. The MDC of 3.5 points for the RMQ represents a minimum of 15% change in the score to exceed measurement error, whereas the MDCs of 2.5 and 2.1 for the ADL and Recreational domains of the LLTQ respectively, are notably lower, representing 6% and 5% change in the measure.

Calculations of the SEM and MDC are largely influenced by the reliability estimate used to calculate these values and when applied clinically, will be affected by baseline physical function where those with scores near the centre of the scale require more change than those with very high or low initial scores (Deyo, Diehr & Patrick, 1991). For this reason, the application of these distribution based statistics to individual patients in clinical settings may be limited. There is also caution needed in comparing these findings to those noted in previous studies, as past reports have varied widely. This is particularly notable in reports of the RMQ, where the MDC has ranged from 2-3 to 8.6 points (Beurskens, de Vet & Koke, 1996; Davidson & Keating, 2002; Jordan, Dunn, Lewis & Croft, 2002; Ostelo, de Vet, Knol & van der Brandt, 2004; Patrick et al, 1995; Stratford et al, 1996). Comparison of SEMs and MDCs across studies may also be limited by varied subject scores, test-retest periods used to calculate reliability and the statistic employed as an estimate of reliability (de Vet et al, 2006). For these reasons, it has been widely recommended that the SEM and MDC is generated specifically for the population and setting under study in addition to the calculation of the MCID (Beaton, Boers & Wells, 2002; de Vet et al, 2006, van der Roer et al, 2006). The findings of this investigation do suggest however that the domains of the LLTQ are at least as responsive at detecting change in scores above levels of error than the RMQ in this sample, and also provide important information concerning the error of both outcome measures when applying them in a practical setting.

5.9 Investigation E; Minimal clinical important difference of the LLTQ; Receiver operating characteristic curve

Generating the minimal magnitude of change of an outcome measure that is deemed important to patients is integral in clinical and research settings of individuals or groups to decipher the relevance of change over time and intervention. It is therefore vital that the MCID of outcome measures be established for appropriate and effective use in detecting meaningful change for outcome measures (Beaton, Bombardier, Katz & Wright, 2001; Copay et al, 2007).

There are several reasons for the different methodological approach used within the current study to ascertain the MCID of the LLTQ compared to that noted in recent studies. As the current study employed a population of acute LBP subjects, it was anticipated and subsequently observed that the vast majority of subjects would exhibit rapid and significant improvement in physical function. As a result, adequate sets of data from subjects reporting no improvement at follow up testing would have been very difficult to ascertain. In addition, as this was the first study to investigate the MCID of the LLTQ in a LBP population, there is no comparable data available from the use of this outcome measure within a similar population. Therefore, ROC curves were developed for the LLTQ domains using data from past research investigating the LLTQ to represent scores of subjects not reporting improvement (McNair et al, 2007).

The high area under the curve of 0.95 generated in the current study for both LLTQ domains demonstrates the similarity of both domains regarding their ability to detect meaningful change by representing the probability of making a correct ranking of an improved or non-improved subject, according to change in outcome measure score (Guyatt et al, 2002, Revicki, Hays, Cella & Sloan, 2008; Ward, Marx & Barry, 2000). These values are higher than those reported by McNair et al (2007) of 0.91 for the ADL domain and 0.88 for the Recreational domain using a subject population with lower limb conditions.

Past reports of areas under the curve generated for the RMQ have varied from 0.69 to 0.94, with most scores over 0.8 (Beurskens, de Vet & Koke, 1996; Frost, Lamb & Stewart-Brown, 2008; Grotle, Brox & Vollestad, 2004; Davidson & Keating, 2006;

Lauridsen et al, 2006; Kuijer et al, 2007; Kovacs et al, 2007; Stratford, Binkley, Riddle & Guyatt, 1998;). Although the area under the curve statistics offer a measure of clinical responsiveness of an outcome measure, there are several limitations in directly comparing these values with past reports from other outcome measures, including the RMQ. Values of the area under the curve of the RMQ within past research have been generated through use of RMQ change scores from subjects reporting at least the minimum level of change, thus including change scores of subjects reporting improvement higher than the minimal level on the GPE scale, for example 'much improved' and 'completely recovered'. The advantage of our approach is that results are representative of score changes of only those subjects reaching the minimal definition of improvement, therefore allowing the assumption to be made that the subsequent MCID calculation is representative of minimal change. The current investigation has demonstrated that although using a more conservative approach of using subject scores where only the minimal level of change was indicated, the areas under the curves generated for the LLTQ domains are very high, indicating its ability to detect meaningful change in this sample.

Contrasting the area under the curves generated for the LLTQ with those of the RMQ from previous research may facilitate the comparison between the current results and research pertaining to the MCID of the RMQ (Figure 5.1). It should be noted however that LLTQ change scores in the current study pertain to data only from subjects reporting improvement at the minimal level of change, and the three RMQ studies used for comparisons have dichotomised results into scores from unchanged subjects or those reporting change at or above the minimal level of improvement. Figure 5.1 demonstrates that despite the more conservative approach used in the current study, the areas under the curves generated for the two domains of the LLTQ are similar. While statistical analysis can be used to assess differences across areas under the curves, it was not possible within the current study. The primary reason was that the current study did not include a sample of unchanged acute LBP patients, rather LLTQ data from an unchanged population with lower limb conditions were used.
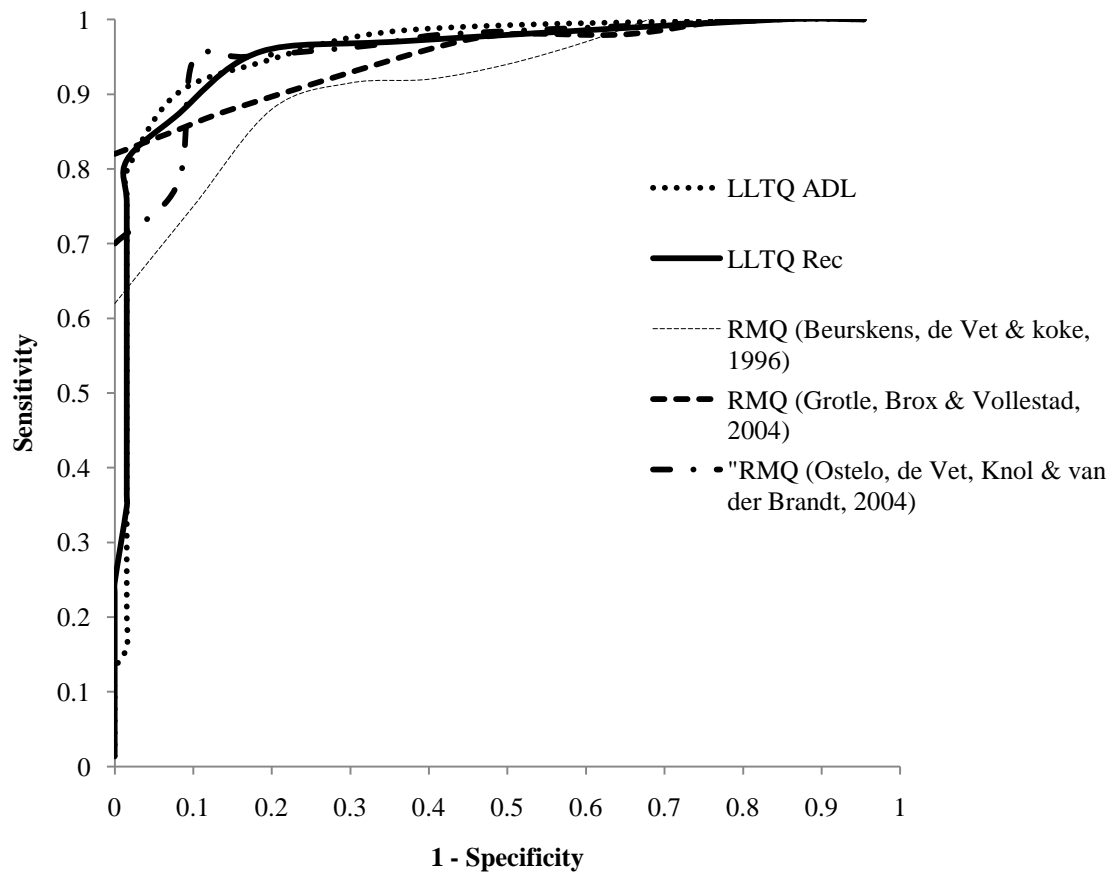
Figure 5.1: Comparison of LLTQ and RMQ ROC Curves.

When standardising the MCID to a percentage of change score, the MCID of 3 points within the current study for the LLTQ domains represents a 7.5% score change which is lower than MCIDs reported for the RMQ in past research (Beurskens, de Vet & Koke, 1996;; Davidson & Keating, 2002; Grotle, Brox & Vollestad, 2004; Kovacs et al, 2007; Jordan, Dunn, Lewis & Croft, 2006; Lauridsen et al, 2006; Ostelo, de Vet, Knol & van der Brandt, 2004; Ostelo et al, 2008; Stratford, Binkley, Riddle & Guyatt, 1998). As discussed within the literature review, the MCID generated for the RMQ using anchor based methods has varied widely, due to the large heterogeneity in study methodology, subjects employed and external criterion used to define meaningful change. However, the current evidence base pertaining to the RMQ has suggested an absolute MCID value

of close to 4-5 points, relating to 16-21% of the measure, a substantially higher requirement of change than has been indicated for the LLTQ domains within the current study.

The MCIDs generated for the LLTQ with a likelihood ratio of more than 10 generated within the current study represent a change in score with the highest corresponding sensitivity and specificity to detect meaningful change using the LLTQ. This may assist in clinical application of the LLTQ, as it has been suggested that accuracy of measures can be interpreted more readily if presented in this manner (Jaeschke, Guyatt & Sackett, 1994) and provide an indication of probability of accuracy in detecting change (Hayden & Brown, 1999).

The secondary investigation regarding the influence of initial scores on MCIDs generated the same value regardless of category used to indicate the level of limitation of function at baseline. The logical way to categorise subjects initial scores was to group them according to what was considered severe (less than 20 points), moderate (20 – 30 points) or minor (over 30 points) limitation in function at initial scoring rather than categorising subjects initial scores based on arbitrary divisions of scores, as was indicated by the founder of the outcome measure (McNair et al, 2007). It was found that there were very few subjects with initial scores over 30 points indicating minor limitation in function at baseline, particularly in the Recreational domain, and subsequently there were insufficient subject scores within this category to generate an MCID through ROC curve analysis. Despite this, MCIDs generated for those considered moderate and severe at baseline were 3 points for both domains of the outcome measure, indicating little influence of initial scores on the MCID of the LLTQ. There is further support for this finding within the results of the tertiary investigation using a regression analysis of initial scores and change scores for the LLTQ. Although results show a negative relationship between initial scores and change scores, the low r squared values of 0.25 and 0.18 for the ADL and Recreational domains respectively demonstrate that initial scores do not play a large role in the variance of change scores. Although limited by only two categories and low subject numbers, these findings may have important clinical implications as they indicate a similar MCID regardless of severity of limitation in function at baseline. This finding is contrary to several studies addressing the RMQ (Kovacs et al, 2007; Demoulin et al, 2010; Stratford, Binkley, Riddle & Guyatt, 1998) where variation in MCIDs have been demonstrated using ROC

analysis across subject groups with varying baseline scores. Whether or not these findings suggest that the LLTQ is not as susceptible to initial scoring as the RMQ cannot be substantiated without further research using the LLTQ with larger subject numbers and further analysis of change scores of groups based on all three or more categories at initial scoring.

5.10 Limitations

This section will outline the limitations of the current study, with particular regard to the outcome measures used, procedures and statistical analyses performed.

5.10.1 Outcome measures and procedures

The use of questionnaires as the primary outcome measures may be influenced by the participant's interpretation, their own reflection of physical function, fear avoidance associated with the injury and potential apprehension about change in treatment (Myers et al, 2008; Lackner, Carosella & Feuerstein, 1996). In addition, responses to items of questionnaires are based on the subjects own self efficacy, therefore subjects with lower self efficacy may generate lower physical functioning scores, particularly when the questionnaires include tasks that have not been attempted (Feuerstein & Beattie, 1995; Woby, Urmston & Watson, 2007).

The effect of patient bias with the use of a GPE scale cannot be excluded as this score has been reportedly associated with recall bias and affected by a correlation with the present state of the individual (Copay et al, 2007; Norman, Stratford & Regehr, 1997). In addition, although the GPE scale employed asks the participant to rate their change in specific regard to physical function, it is impossible to exclude the subject's consideration of change experienced in other areas such as pain, quality of life or social and emotional functioning. It also cannot be assumed that the perception of the level of 'improved' on the GPE scale represents meaningful change in limitation of physical function for all individuals. The reliability of the GPE scale was also not assessed within the current study.

5.10.2 Statistical Analysis

In order to generate the MCID of an outcome measure, ROC curves are generated based on scores from patients reporting meaningful change and those unchanged (Revicki, Hays, Cella & Sloan, 2008). An unchanged population was not tested within the current study, rather data was used from subjects reporting no change when using the LLTQ in a population with lower limb conditions within a previous study (McNair et al, 2007). Using data from a sample of unchanged subjects from a different population to the one employed within the current study may not give an accurate representation of the variability of change scores in subjects in the population of interest, potentially influencing the resultant MCID. In addition, as scores of the RMQ in a LBP population reporting no change were not available, the MCID of the RMQ was not generated using anchor based methods. This resulted in the MCID generated for the LLTQ being contrasted to that of the RMQ based only on past reports.

In order to calculate the MCID of the LLTQ, change scores were gathered from subjects who reported a level of 'improved' on the GPE scale when retested at each follow up visit for physiotherapy treatment. This method resulted in LLTQ change scores representing minimum change being gathered from only 50 of the total 69 subjects included within the study. Higher subject numbers for this 'improved' only group may potentially have been achieved if more frequent retesting was used to capture subjects at this level of change, or higher subject numbers were initially recruited.

CHAPTER 6: SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.1 Summary and Conclusion

The LLTQ is an outcome measure originally designed for use in populations with lower limb conditions. Although not investigated in a LBP population, tasks assessed within the measure, when contrasted to the WHO ICF model, are consistent with those reported as important in the assessment of individuals with LBP (Cieza et al, 2004; McNair et al, 2007; Stier-Jarmer, Cieza, Borchers & Stucki, 2009). The measure offers delineation of tasks relating to ADLs and recreational functioning across a five point likert scale, assessing both the ability to achieve tasks and the importance of each task, thereby assisting application of the measure across different populations and stages of rehabilitation whilst recognising the priorities of the individual. A measure that is also practical, valid and responsive, as demonstrated with the LLTQ, would address some of the burden on clinicians in employing outcome measures in practice, and may facilitate research through increasing subject numbers, pooling of results and increasing significance within trials.

Sixty-nine acute LBP patients were recruited prior to the commencement of physiotherapy treatment and completed the LLTQ and RMQ. Subjects repeated completion of the measures again when a level of at least 'improved' on the GPE scale was indicated at subsequent treatment sessions and at discharge from treatment. The RMQ and LLTQ demonstrated significant score changes between baseline and 'improved', and mean score changes were found to vary according to baseline scores with a significant but small negative correlation between initial scores and change scores. Outcome measure data were then analysed to investigate the following key properties;

Investigation A; Content validity was investigated through analysis of the importance ratings of tasks of the LLTQ in addition to a comparison with the ICF core sets for LBP model. Results highlighted the priorities of both ADLs and recreational activities within the subject group, consistent with the ICF model. Relating tasks to the ICF model also identified areas addressed by the LLTQ in relation to social and sporting activities, not included within other commonly used outcome measures.

Investigation B; Construct validity for the LLTQ was investigated through correlation analysis of scores with the RMQ. Results demonstrated a moderate correlation between the measures, supporting the construct validity of the LLTQ in relation to assessing limitation in physical function in a LBP population.

Investigation C; Statistical responsiveness of the LLTQ was demonstrated by high effect size's and SRM's of over 0.8.

Investigation D; The level of error associated with the score was investigated with an MDC of less than 3 points for both domains of the LLTQ.

Investigation E; The MCID of the two domains of the LLTQ were estimated via anchor based analysis, generating ROC curves to represent optimum sensitivity and specificity of the measure. This method estimated a cut off point of 3 points to represent the MCID, or a percentage score increase of 7.5%. This was contrasted to the RMQ, where a review of the literature indicated an MCID of 4-5 RMQ points or score reduction of approximately 30-40%. Repeated analysis of MCIDs based on subject initial LLTQ scores of 0-19 and 20-30 points generated the same MCID of 3 points for both domains regardless of initial score. Correlation analysis of initial LLTQ scores and score changes also demonstrated that although a negative relationship was present, this correlation was not strong, indicating a consistency in MCID of the LLTQ across baseline scores.

Results of this study set important benchmarks regarding the psychometric properties of the LLTQ within an acute LBP population for measuring limitation in physical function. Through contrasting the measure against a widely recognised and employed LBP outcome measure, the LLTQ has demonstrated several advantages relating to the content and construct validity of the measure. The LLTQ has demonstrated a sound ability to detect statistical and clinically meaningful change within this sample, promoting its use in populations with acute LBP and lower limb conditions within both clinical and research settings.

6.2 Recommendations for future research

As this is the first study to investigate the use of the LLTQ in a LBP population, there are several questions remaining regarding the use of the measure. Based on the review of literature and the results of the study, there are three key areas identified as avenues for future research.

The results of the current study highlight the priorities of individuals regarding functional tasks through LLTQ importance scores. Investigation into the most appropriate method of weighting tasks according to these importance scores may result in an increased ability of the LLTQ to detect meaningful change that reflects the functional priorities of the individual or clinical population assessed.

Further exploration regarding the influence of baseline scores on the MCID of the LLTQ may be warranted using larger subgroups according to specific baseline score categories. This may identify specific cut off points to represent the MCID for particular scores at baseline thereby assisting clinical use of the measure.

Research extending from the current study may also include examining the use of the LLTQ in other populations. With the LLTQ demonstrating good validity and responsiveness in sample populations with lower limb conditions (McNair et al, 2007) and acute LBP in the current study, use in further clinical groups such as chronic LBP may be warranted following further investigation.

REFERENCES

Abrams, D., Davidson, M., Harrick, J., Harcourt, P., Zylinski, M., & Clancy, J. (2006). Monitoring the change: Current trends in outcome measure usage in physiotherapy. *Manual Therapy, 11*(1), 46-53.

Accident Compensation Corporation. (2007). Specific back injury—how big is the problem? Back Injury Prevention. Wellington

ACORPA, Inc. (2006). Australian Standards for Physiotherapy; safe and effective practice. Fortitude Valley, QLD: Australian Physiotherapy Council.

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall.

Anagnostis, C. P., Gatchel, R. J. P., & Mayer, T. G. M. D. (2004). The pain disability questionnaire: a new psychometrically sound measure for chronic musculoskeletal disorders. *Spine, 29*(20), 2290-2302.

Andersson, G.B. (1999). Epidemiological features of chronic low-back pain. Lancet, 14, 354(9178):581-5.

Aseltine, R. H., Carlson, K. J., Fowler, F. J., & Barry, M. J. (1995). Comparing prospective and retrospective measures of treatment outcomes. *Medical Care, 33*(4), S67-S77.

American Physical Therapy Association. (2001). Guide to physical therapist practice. *Physical Therapy, 81*(1), 9-746.

Barber-Westin, S. D., Noyes, F. R., & McCloskey, J. W. (1999). Rigorous statistical reliability, validity, and responsiveness testing of the cincinnati knee rating system in 350 subjects with uninjured, injured, or anterior cruciate ligament-reconstructed knees. *The American Journal of Sports Medicine, 27*(4), 402-416.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215-220.

Bassett, S. F., & Prapavessis, H. (2007). Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains. *Physical Therapy, 87*(9), 1132-1143.

Beaton, D. E. (2000). Understanding the relevance of measured change through studies of responsiveness. *Spine, 25*, 3192 - 3199.

Beaton, D. E., Boers, M., & Wells, G. A. (2002). Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current Opinion in Rheumatology, 14*(2), 109-114.

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology, 54*, 1204 - 1217.

Beattie, P., & Maher, C. (1997). The role of functional status questionnaires for low back pain. *Australian Physiotherapy, 43*(1), 29-39.

Beurskens, A. J., de Vet, H. C., & Koke, A. J. (1996). Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain, 65*, 71 - 76.

Binkley, F. M., & Stratford, P. (1999). Applying the results of self-report measures to individual patients: an example using the Roland-Morris questionnaire *Journal of Orthopaedic & Sports Physical Therapy, 29*(4), 232-239.

Bland, J. M., & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine, 20*(5), 337-340.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*, 135-160.

Bombardier, C., Hayden, J., & Beaton, D. E. (2001). Minimal clinically important difference. Low back pain: outcome measures. *The Journal of Rheumatology, 28*(2), 431-438.

Bombardier, C. M. D. F. (2000). Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine, 25*(24), 3100-3103.

Carey, T. S., Garrett, J., Jackman, A., McLaughlin, C., Fryer, J., Smucker, D. R. (1995). The outcomes and costs of care for acute low back pain among patients seen by primary care practitioners, chiropractors, and orthopedic surgeons. *New England Journal of Medicine, 333*(14), 913-917.

Caulfield, B., & Reilly, C. (2003). Outcome measurement in physiotherapy—why and how? *Physiotherapy Ireland, 24*(1), 8-11.

Cella, D., Eton, D. T., Lai, J.-S., Peterman, A. H., & Merkel, D. E. (2002). Combining anchor and distribution-based methods to derive minimal clinically important differences on the functional assessment of cancer therapy (fact) anemia and fatigue scales. *Journal of Pain and Symptom Management, 24*(6), 547-561.

Chansirinukor, W., Maher, C. G., Latimer, J., & Hush, J. (2004). Comparison of the functional rating index and the 18-item Roland-Morris Disability Questionnaire: responsiveness and reliability. *Spine, 30*, 141 - 145.

Chatman, A., Hyams, S., Neel, J., Binkley, J.M., Stratford, P.W., Schomberg, A. & Stabler, M. (1997).The Patient-Specific Functional Scale: measurement properties in patients with knee dysfunction, *Physical Therapy* 77, 820–829.

Childs, J. D, Piva, S. R., & Fritz, J. M. (2005). Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine, 30*(11), 1331-1334.

Cieza, A., Ewert, T., Ammann, T. B., Chatterji, S., Kostanjsek, N., & Stucki, G. (2004). Development of ICF Core Sets for patients with chronic conditions. *Journal of Rehabilitation Medicine, 36*(4 supp 44), 9 - 11.

Cieza, A., & Stucki, G. (2005). Content comparison of health-related quality of life (HRQOL) instruments based on the international classification of functioning, disability and health (ICF). *Quality of Life Research, 14*(5), 1225-1237.

Clark, L. A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

Cochrane, W. G. (1977). *Sampling Techniques* (3rd. ed.). London: Wiley.

Cohen, J. (1977). *Statistical Power: analysis for the behavioural sciences. .* New York Academic Press

Cook, C., Cleland, J., & Huijbregts, M. P. J. (2007). Creation and critique of studies of diagnostic accuracy: use of the stard and quadas methodological quality assessment tools. *Journal of Manual and Manipulative Therapy, 15*(2), 93-102.

Copay, A. G., Subach, B. R., Glassman, S. D., Polly Jr, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal, 7*(5), 541-546.

Copeland, J. M., Taylor, W. J., & Dean, S. G. (2008). Factors influencing the use of outcome measures for patients with low back pain: a survey of new zealand physical therapists. *Physical Therapy, 88*(12), 1492-1505.

Cornell, J.A. & Berger, R.D. (1986). Factors that influence the value of the coefficient of determination in simple linear and nonlinear regression models. *Phytopathology, 77,* 63 – 70

Costa, L. O. P., Maher, C. G., Latimer, J., Ferreira, P. H., Pozzi, G. C. & Robiero, R.N. (2007). Psychometric characteristics of the Brazilian-Portuguese versions of the Functional Rating Index and the Roland Morris Disability Questionnaire. Spine, 32 (17), 1902-1907

Costa, L. O. P., Maher, C. G., Latimer, J., Ferreira, P. H., Ferreira, M. L. & Pozzi, G. C. (2008). Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil: which one is the best? *Spine, 33*(22), 2459-2463.

Costa, L. O., Maher, C. G., & Latimer, J. P. (2007). Self-report outcome measures for low back pain: searching for international cross-cultural adaptations. *Spine, 32*(9), 1028-1037.

Coste, J., Delecoeuillerie, G., de Lara, A. C., LeParc, J. M., & Paolaggi, J. B. (1994). Clinical course and prognostic factors in acute low back pain: an inception cohort study in primary care practice. *British Medical Journal, 308*(6928), 577-580.

Croft, P. R., Macfarlane, G. J., Papageorgiou, A. C., Thomas, E., & Silman, A. J. (1998). Outcome of low back pain in general practice: a prospective study. *British Medical Journal, 316*(7141), 1356-1359.

Crombez, G., Vlaeyen, J. W. S., Heuts, P. H. T. G., & Lysens, R. (1999). Pain-related fear is more disabling than pain itself: evidence on the role of pain-related fear in chronic back pain disability. *Pain, 80*(1-2), 329-339.

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology, 56*, 395 – 407.

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2004). An integrated method to determine meaningful changes in health-related quality of life. *Journal of Clinical Epidemiology, 57*(11), 1153-1160.

Davidson, M., & Keating, J. L. (2002). A comparison of five low back disability questionnaires: reliability and responsiveness. *Physical Therapy, 82*, 8 - 24.

de Bruin, A. F., Diederiks, J. P. M., de Witte, L. P., Stevens, F. C. J., & Philipsen, H. (1997). Assessing the responsiveness of a functional status measure: The sickness impact profile versus the SIP68. *Journal of Clinical Epidemiology, 50*(5), 529-540.

de Vet, H., Terwee, C., Ostelo, R., Beckerman, H., Knol, D., & Bouter, L. (2006). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes, 4*(1), 54.

de Vet, H. C., Ostelo, R. W., Terwee, C. B., van der, R. N., Knol, D. L., & Beckerman, H. (2007). Minimally important change determined by a visual method intergrating an anchor-based and distribution-based approach. . *Quality of Life Research, 16*, 131-142.

de Vet, H. C. W., Terwee, C. B., & Bouter, L. M. (2003). Current challenges in clinimetrics. *Journal of Clinical Epidemiology, 56*(12), 1137-1141.

Demoulin, C., Ostelo, R., Knottnerus, J. A., & Smeets, R. J. E. M. (2010). What factors influence the measurement properties of the Roland-Morris disability questionnaire? *European Journal of Pain, 14*(2), 200 - 206.

Dettori, J., Bullock, S., Sutlive, T., Franklin, R., & Patience, T. (1995). The effects of spinal flexion and extension exercises and their associated postures in patients with acute low back pain. *Spine, 20*(21).

Deyo, R. A., & Carter, W. B. (1992). Strategies for improving and expanding the application of health status measures in clinical settings: a researcher-developer viewpoint *Medical Care, 30*(5), 176-186.

Deyo, R. A., & Centor, R. M. (1986). Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of Chronic Disease, 39*, 897 - 906.

Deyo, R. A., Diehr, P., & Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures statistics and strategies for evaluation. *Controlled Clinical Trials, 12*(4, Supplement 1), S142-S158.

Deyo, R. A., Battie, M. P., Beurskens, A. T., Bombardier, C. M., Croft, P. M., Koes, B. P., et al. (1998). Outcome measures for low back pain research: a proposal for standardized use. *Spine, 23*(18), 2003-2013.

Diehr, P., Chen, L., Patrick, D., Feng, Z., & Yasui, Y. (2005). Reliability, effect size, and responsiveness of health status measures in the design of randomized and cluster-randomized trials. *Contemporary Clinical Trials, 26*(1), 45-58.

Dionne, C. E., Koepsell, T. D., Von Korff, M., Deyo, R. A., Barlow, W. E., & Checkoway, H. (1997). Predicting long-term functional limitations among back pain patients in primary care settings. *Journal of Clinical Epidemiology, 50*(1), 31-43.

Domholdt, E. (2005). *Rehabilitation research; principles and applications* (3rd ed.). St Louis: Elsevier Saunders.

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain, 113*(1-2), 9-19.

Ehrlich, G. E. (2003). Back pain. *The Journal of Rheumatology, 67*, 26-31.

Fairbanks, J., Davies, J., Couper, J., & O'Brien, J. (1980). The Oswestry low back pain disability questionnaire. *Physiotherapy, 66*, 271-272.

Farrar, J. T., Portenoy, R. K., Berlin, J. A., Kinman, J. L., & Strom, B. L. (2000). Defining the clinically important difference in pain outcome measures. *Pain, 88*(3), 287-294.

Farrar, J. T., Young, J. P., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain, 94*(2), 149-158.

Ferguson, R. J., Robinson, A. B., & Splaine, M. (2002). Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. *Quality of Life Research, 11*(6), 509-516.

Ferreira, M. L., Ferreira, P. H., Latimer, J., Herbert, R. D., Hodges, P. W., Jennings, M. D., et al. (2007). Comparison of general exercise, motor control exercise and spinal manipulative therapy for chronic low back pain: A randomized trial. *Pain, 131*(1-2), 31-37.

Feuerstein, M., & Beattie, P. (1995). Biobehavioural factors affecting pain and disability in low back pain; mechanisms and management. *Clinical Perspectives, 75*(4), 267-280.

Fischer, D., Stewart, A. L., Bloch, D. A., Lorig, K., Laurent, D., & Holman, H. (1999). Capturing the patient's view of change as a clinical outcome measure. *Journal of the American Medical Association, 282*, 1157 - 1162.

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D., & Cox, D. (1992). Quality of life measures in health care. I: Applications and issues in assessment. *British Medical Journal, 305*(6861), 1074-1077.

Fritz, J. M., George, S. Z., & Delitto, A. (2001). The role of fear-avoidance beliefs in acute low back pain: relationships with current and future disability and work status. *Pain, 94*(1), 7-15.

Fritz, J. M.., Delitto, A. P., & Erhard, R. E.. (2003). Comparison of classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain: a randomized clinical trial. *Spine, 28*(13), 1363-1371.

Fritz, J. M., Hebert, J. D., Koppenhaver, S. P., & Parent, E. P. (2009). Beyond minimally important change: defining a successful outcome of physical therapy for patients with low back pain. *Spine, 34*(25), 2803-2809.

Frost, H., Lamb, S., & Stewart-Brown, S. (2008). Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index and the Roland Morris Disability Questionnaire for patients with subacute and chronic LBP. *Spine, 33*(22), 2450-2457.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. A., & Hays, R. D. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health, 10*(s2), S94-S105.

Garratt, A. M. P., Klaber Moffett, J. P. M., & Farrin, A. J. M. (2001). Responsiveness of generic and specific measures of health outcome in low back pain. *Spine, 26*(1), 71-77.

Gilson, B.S., Gilson, J.S., Bergner, M., Bobbit, R.A., Kressel, S., Pollard, W.E. & Vesselago, M. (1975). The sickness impact profile; development of an outcome measure of health care. American Journal of Public Health, 65 (12), 1304-1310

Greenhalgh, J., Long, A. F., & Flynn, R. (2005). The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Social Science & Medicine, 60*(4), 833-843.

Greenhalgh, J., & Meadows, K. (1999). The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *Journal of Evaluation in Clinical Practice, 5*(4), 401-416.

Greenough, C. G., & Fraser, R. D. (1992). Assessment of outcome in patients with low-back pain. *Spine, 17*(1), 36.

Grimmer, K., Sheppard, L., Pitt, M., Magarey, M., & Trott, P. (1999). Differences in stakeholder expectations in the outcome of physiotherapy management of acute low back pain. *International Journal of Quality Health Care, 11*(2), 155-162.

Grotle, M., Brox, J. I., Glomsrød, B., Lønn, J. H., & Vøllestad, N. K. (2007). Prognostic factors in first-time care seekers due to acute low back pain. *European Journal of Pain, 11*(3), 290-298.

Grotle, M., Brox, J. I., & Vollestad, N. K. (2004). Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. *Spine, 29*, E492 - E501.

Grotle, M., Brox, J. I., & Vollestad, N. K. (2005). Functional Status and disability questionnaires: what do they assess?: a systematic review of back-specific outcome questionnaires. *Spine, 30*, 130 - 140.

Grotle, M., Vøllestad, N. K., Veierød, M. B., & Brox, J. I. (2004). Fear-avoidance beliefs and distress in relation to disability in acute and chronic low back pain. *Pain, 112*(3), 343-352.

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., Norman, G. R., & Clinical Significance Consensus Meeting, G. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clin. Proc., 77*(4), 371-383.

Hägg, O., Fritzell, P., & Nordwall, A. (2003). The clinical importance of changes in outcome scores after treatment for chronic low back pain. *European Spine Journal, 12*(1), 12-20.

Hagg, O., Fritzell, P., Oden, A., & Nordwall, A. (2002). Simplifying outcome measurement: Evaluation of instruments for measuring outcome after fusion surgery for chronic low back pain. *Spine, 27*(11), 1213-1222.

Haigh, R., Tennant, A., Biering-Sorensen, F., Grimby, G., Marincek, C., Phillips, S., Ring, H., Tesio, L. & Thonnard, J.L. (2001).The use of outcome measures in physical medicine and rehabilitation within Europe. *Journal or Rehabilitation Medicine*, 33 (6), 273 – 278

Hanley, R. & McNeill, T. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Radiology* 143(1)29-36 1982

Hare-Mortensen, L., Lauridsen, H., & Grunnet-Nilsson, N. (2006). The Relative Responsiveness of 3 Different Types of Clinical Outcome Measures on Chiropractic Patients with Low Back Pain. *Journal of Manipulative and Physiological Therapeutics, 29*(2), 95-99.

Hayden, J. A., van Tulder, M. W., Malmivaara, A. V., & Koes, B. W. (2005). Meta-analysis: exercise therapy for nonspecific low back pain. *Annals of Internal Medicine, 142*(9), 765-775.

Hayden, S. R., & Brown, M. D. (1999). Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. *Annals of Emergency Medicine, 33*(5), 575-580.

Hebert, R., Spiegelhalter, D.J. & Brayne, C. (1997). Setting the minimal metrically detectable change on disability rating scales. *Archives of Physical Medicine and Rehabilitation, 78,* 1305 - 1308

Hevey, D., & McGee, H. (1998). The effect size statistic: useful in health outcomes research? . *Journal of Health Psychology, 3*, 163-170.

Hides, J., Gilmore, C., Stanton, W., & Bohlscheid, E. (2008). Multifidus size and symmetry among chronic LBP and healthy asymptomatic subjects. *Manual Therapy, 13*(1), 43-49.

Hides, J. A. P., Richardson, C. A. P., & Jull, G. A. M. (1996). Multifidus muscle recovery is not automatic after resolution of acute, first-episode low back pain. *Spine, 21*(23), 2763-2769.

Hides, J. A., Jull, G. A., & Richardson, C. A. (2001). Long-term effects of specific stabilizing exercises for first-episode low back pain. *Spine, 26*(11), e243-e248.

Higginson, I. J., & Carr, A. J. (2001). Using quality of life measures in the clinical setting. *British Medical Journal, 322*, 1297–1300.

Hollingworth, W., Medina, L. S., Lenkinski, R. E., Shibata, D. K., Bernal, B., Zurakowski, D., et al. (2006). Interrater reliability in assessing quality of diagnostic accuracy studies using the quadas tool: a preliminary assessment. *Academic Radiology, 13*(7), 803-810.

Horner, D., & Larmer, P. (2006). Health outcome measures. *New Zealand Journal of Physiotherapy, 34*(1), 17-24.

Huijbregts, M. P. J., Myers, A. M., Kay, T. M., & Gavin, T. S. (2002). Systematic outcome measurement in clinical practice: challenges experienced by physiotherapists. *Physiotherapy Canada, 54*(1), 25

Hurst, N. P., Kind, P., Ruta, D., Hunter, M., & Stubbings, A. (1997). Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Rheumatology, 36*(5), 551-559.

Jacob, T., Baras, M., Zeev, A., & Epstein, L. (2001). Low back pain: Reliability of a set of pain measurement tools. *Archives of Physical Medicine and Rehabilitation, 82*(6), 735-742.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12.

Jaeschke, R., Guyatt, G., & Sackett, D. L. (1994). Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *The Journal of the Americal Medical Association, 271*(5), 389-439.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials, 10*, 407 - 415.

Jette, A. M. (2006). Toward a Common Language for Function, Disability, and Health. *Physical Therapy, 86*(5), 726-734.

Jette, A. M., Assmann, S. F., Rooks, D., Harris, B. A., & Crawford, S. (1998). Interrelationships among disablement concepts. *The Journals of Gerontology: Series A Biological sciences and medical sciences, 53*, 395-404.

Jette, D. U., & Jette, A. M. (1996). Physical therapy and health outcomes in patients with spinal impairments. *Physical Therapy, 76*, 930 - 941.

Johansson, E., & Lindberg, P. (1998). Subacute and chronic low back pain. Reliability and validity of a Swedish version of the Roland and Morris Disability Questionnaire. *Scandinavian Journal of Rehabilitation Medicine, 30*(3), 139-143.

Jordan, K., Dunn, K. M., Lewis, M., & Croft, P. (2006). A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. *Journal of Clinical Epidemiology, 59*(1), 45-52.

Juni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *British Medical Journal, 323*(7303), 42-46.

Kamper, S. J., Maher, C. G., & Mackay, G. (2009). Global rating of change scales: a review of strengths and weaknesses and considerations for design. *Journal of Manual & Manipulative Therapy, 17*(3), 163-170.

Katz, J. N., Larson, M. G., Phillips, L. B., Fossel, A. H., & Liang, M. H. (1992). Comparative measurement sensitivity of short and longer health status instruments. *Medical Care, 30*(10), 917-925.

Kay, T. M., Myers, A. M., & Huijbregts, M. P. J. (2001). How far have we come since 1992? A comparative survey of physiotherapists' use of outcome measures. *Physiotherapy Canada, 53*(4), 268-275.

Kazis, L. E., Anderson, J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. *Medical Care, 27*(3), 178-189.

Kent, P., & Keating, J. L. (2005). Classification in nonspecific low back pain: what methods do primary care clinicians currently use? *Spine, 30*(12), 1433-1440

Khorsan, R., Coulter, I. D., Hawk, C., & Choate, C. G. (2008). Measures in chiropractic research: choosing patient-based outcome assessments. *Journal of Manipulative and Physiological Therapeutics, 31*(5), 355-375.

Kirkness, C., & Korner-Bitenshky, N. (2002). Prevalence of outcome measure use by physiotherapists. . *Physiotherapy Canada, 54*, 249-257.

Kitaoka, H.I., Adellar, A.R., Nunley, J.A., Myerson, M.S. & Sanders, M. (1994). Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes, *Foot Ankle 15,* 349–353.

Koes, B.W., van Tulder, M,W. & Thomas, S. (2006). Diagnosis and treatment of low back pain. British Medical Journal, 332, 1430 - 1434

Kopec, J. A., Esdaile, J. M., Abrahamowicz, M., Abenhaim, L., Wood-Dauphinee, S., Lamping, D. L., et al. (1995). The Quebec back pain disability scale: measurement properties. *Spine, 20*(3), 341-352.

Kovacs, F. M., Abraira, V. P., Royuela, A. M., Corcoll, J. M. D., Alegre, L. M., Cano, A. M. (2007). Minimal Clinically Important Change for Pain Intensity and Disability in Patients With Nonspecific Low Back Pain. *Spine, 32*(25), 2915-2920.

Kovacs, F.M., Abraira, V., Zamora, J., Teresa Gil del Real, M., Llobera, J. & Fernández, C. (2004). Correlation between pain, disability, and quality of life in patients with common low back pain. Spine, 29 (2), 206 - 210

Kucukdeveci, A. A., Tennant, A., Elhan, A. H., & Niyazoglu, H. (2001). Validation of the turkish version of the roland-morris disability questionnaire for use in low back pain. *Spine, 26*(24), 2738-2743.

Kuijer, W., Brouwer, S., Dijkstra, P. U., Jorritsma, W., Groothoff, J. W., & Geertzen, J. H. B. (2005). Responsiveness of the Roland Morris Disability Questionnaire: consequences of using different external criteria. *Clinical Rehabilitation, 19*(5), 488-495.

Lackner, J. M., Carosella, A. M., & Feuerstein, M. (1996). Pain expectancies, pain and functional self efficacy expectancies as determinants of disability in patients with chronic low back disorders. *Journal of Consulting and Clinical Psychology, 64*(1), 212-220.

Lassere, M.N., van der Heijde, D., Johnson, K., Bruynesteyn, K., Boonen, A., Verhoeven, A., Emery, P. & Boers, M. (2001). Robustness and generalizability of smallest detectable difference in radiological progression. *The Journal of Rheumatology, 2* 8(4):911-3.

Lauridsen, H., Hartvigsen, J., Manniche, C., Korsholm, L., & Grunnet-Nilsson, N. (2006). Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskeletal Disorders, 7*(1), 82.

Lauridsen, H. H., Hartvigsen, J., Korsholm, L., Grunnet-Nilsson, N., & Manniche, C. (2007). Choice of external criteria in back pain research: Does it matter? Recommendations based on analysis of responsiveness. *Pain, 131*(1-2), 112-120.

Leclaire, R. M. D., Blier, F. M. D., Fortin, L. M. D., & Proulx, R. M. D. (1997). A cross-sectional study comparing the oswestry and roland-morris functional disability scales in two populations of patients with low back pain of different levels of severity. *Spine, 22*(1), 68-71.

Leidy, N. K., & Wyrwich, K. W. (2004). Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). *COPD: Journal of Chronic Obstructive Pulmonary Disease, 2*(1), 157 - 165.

Liddle, S. D., Baxter, D. G., & Gracey, J. H. (2009). Physiotherapists' use of advice and exercise for the management of chronic low back pain: A national survey. *Manual Therapy, 14*(2), 189-196.

Linton, S. J. (2000). A review of psychological risk factors in back and neck pain. *Spine, 25*(9), 1148-1156.

Little, D. G., & Macdonald, D. (1994). The use of the percentage change in oswestry disability index score as an outcome measure in lumbar spinal surgery. *Spine, 19*(19), 2139 - 2143.

Lydick, E., & Epstein, R. S. (1993). Interpretation of quality of life changes. *Quality of Life Research, 2*(3), 221-226.

Lysholm, J., & Gillquist, J. (1982). Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale. *The American Journal of Sports Medicine, 10*(3), 150-154.

Manchikanti, L. (2000). Epidemiology of low back pain. *Pain Physician, 3* (2), 167 - 192

Mannion, A. F., Junge, A., Taimela, S., Mantener, M., Lorenzo, K. S., & Dvorak, J. (2001). Active therapy for chronic low back pain: part 3. factors influencing self-rated disability and its change following therapy. *Spine, 26*(8), 920-929.

Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: a structured review. *Journal of Evaluation in Clinical Practice, 12*(5), 559-568.

May, S. (2003). An outcome audit for musculoskeletal patients in primary care. *Physiotherapy Theory & Practice, 19*(4), 189-198.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. . *Psychological Methods, 1*(1), 30-46.

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research, 4*(4), 293-307.

McNair, P. J., Prapavessis, H., Collier, J., Bassett, S., Bryant, A., & Larmer, P. (2007). The Lower-Limb tasks questionnaire: an assessment of validity, reliability, responsiveness, and minimal important Differences. *Archives of Physical Medicine and Rehabilitation, 88*(8), 993-1001.

Meads, C., & Davenport, C. (2009). Quality assessment of diagnostic before-after studies: development of methodology in the context of a systematic review. *BMC Medical Research Methodology, 9*(1), 3.

Monk, C. (2006). Measurement of the functional improvement in patients receiving physiotherapy for musculoskeletal conditions. *New Zealand Journal of Physiotherapy, 34*(2), 50-55.

Myers, S., Phillips, R., Davis, R., Cherkin, D., Legedza, A., Kaptchuk, T. (2008). Patient expectations as predictors of outcome in patients with acute low back pain. *Journal of General Internal Medicine, 23*(2), 148-153.

Nagi, S. Z. (1964). A study in the evaluation of disability and rehabilitation potential: concepts, methods, and procedures. *American Journal of Public Health Nations Health, 54*, 1568–1579.

New Zealand Physiotherapy Board (1999). *Registrations Requirements; Competencies and Learning Objectives* (2nd ed.). Wellington, New Zealand.

Nordin, M.L., Skovron, R.H., Weiser, s., Brisson, P., Campello, M., Harwood, K., Crane, M. & Lewis, S. (1997). Early predictors of delayed return to work in patients with low back pain, *Journal of Musculoskeletal Pain*, 5 (2), 5-27

Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies. *Quality of Life Research*, 12(3), 239-249.

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

Norman, G. R. P., Sloan, J. A. P., & Wyrwich, K. W. P. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care, 41*(5), 582-592.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ostelo, R. W., Deyo, R. A., Stratford, P., Waddell, G., Croft, P., Von Korff, M. (2008). Interpreting change scores for pain and functional status in low back pain; towards international consensus regarding minimal important change. *Spine, 33*(1), 90-94.

Ostelo, R. W. J. G., & de Vet, H. C. W. (2005). Clinically important outcomes in low back pain. *Best Practice & Research Clinical Rheumatology, 19*(4), 593-607.

Ostelo, R. W. J. G., de Vet, H. C. W., Knol, D. L., & van den Brandt, P. A. (2004). 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *Journal of Clinical Epidemiology, 57*(3), 268-276.

Patrick, D.L. & Deyo, R.A (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care, 27*:5217-5232, 1989.

Patrick, D. L., Deyo, R. A., Atlas, S. J., Singer, D. E., Chapin, A., & Keller, R. B. (1995). Assessing health-related quality of life in patients with Sciatica. *Spine, 20*(17), 1899-1908.

Pengel, L. H. M., Herbert, R. D., Maher, C. G., & Refshauge, K. M. (2003). Acute low back pain: systematic review of its prognosis. *British Medical Journal, 327*(7410), 323

Pengel, L. H. M. M., Refshauge, K. M. P., & Maher, C. G. P. (2004). Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. *Spine, 29*(8), 879-883.

Redelmeier, D.A. Guyatt, G.H. & Goldstein, R.S. (1996). Assessing the minimal important difference in symptoms; a comparison of two techniques. *Journal of Clinical Epidemiology, 49*, 1215 - 1219

Resnik, L., & Dobrzykowski, E. (2003). Guide to outcomes measurement for patients with low back pain syndromes. *Journal of Orthopaedic & Sports Physical Therapy, 33*(6), 307-318.

Resnik, L., & Dobrzykowski, E. (2005). Outcomes measurement for patients with low back pain. *Orthopaedic Nursing, 24*(1), 14-24.

Revicki, D., Cella, D., Hays, R., Sloan, J., Lenderking, W., & Aaronson, N. (2006). Responsiveness and minimal important differences for patient reported outcomes. *Health and Quality of Life Outcomes, 4*(1), 70.

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102-109.

Rocchi, E., Sisti D, Benedetti P, Valentini M, Bellagamba S, Federici A. (2005).Critical comparison of nine different self-administered questionnaires for the evaluation of disability caused by low back pain. *Europa Medicophysica, 41* (4), 275-281

Roebroeck, M.E., Harlaar, J. & Lankhorst, G.J. (1993). The application of generalisability theory to reliability assessment; an illustration using isometric force measurements. Physical Therapy, 73, 6. 386 - 395

Roland, M., & Fairbank, J. (2000). The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. *Spine, 25*(24), 3115-3124.

Roland, M., & Morris, R. (1983). A study of the natural history of back pain, part 1: development of a reliable and sensitive measure of disability in low back pain. *Spine, 8*, 141–144.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*(2), 341-357.

Rothstein, J. M., & Echternach, J. L. (1993). Primer on measurement; an introductory guide to measurement issues. In A. P. T. Association (Ed.). Alexandria.

Sieben, J. M., Vlaeyen, J. W. S., Tuerlinckx, S., & Portegijs, P. J. M. (2002). Pain-related fear in acute low back pain: the first two weeks of a new episode. *European Journal of Pain, 6*(3), 229-237.

Sigl, T. M. D., Cieza, A. P., Brockow, T. M. D., Chatterji, S. M. D., Kostanjsek, N. M. S., & Stucki, G. M. D. M. S. (2006). Content comparison of low back pain-specific measures based on the international classification of functioning, disability and health (ICF). *Clinical Journal of Pain, 22*(2), 147-153.

Simmonds, M. J. P., Olson, S. L. P., Jones, S. M. D., Hussein, T. M. S., Lee, C. E. B., Novy, D. P., et al. (1998). Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine, 23*(22), 2412-2421.

Stier-Jarmer, M. P., Cieza, A. P., Borchers, M. M. D., & Stucki, G. M. D. M. S. (2009). How to apply the ICF and ICF core sets for low back pain. *Clinical Journal of Pain, 25*(1), 29-38.

Stratford, P., Finch, E., Solomon, P., Binkley, F. M., Gill, C., & Moreland, J. (1996). Using the Roland-Morris Questionnaire to make decisions about individual patients. *Physiotherapy Canada, 48*(2), 107-118.

Stratford, P. W., Binkley, J., Solomon, P., Finch, E., Gill, C., & Moreland, J. (1996). Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Physical Therapy, 76*(4), 359-365.

Stratford, P. W., Binkley, J., Solomon, P., Gill, C., & Finch, E. (1994). Assessing change over time in patients with low back pain. *Physical Therapy, 74*(6), 528-533.

Stratford, P. W., Binkley, J. M., Riddle, D. L., & Guyatt, G. H. (1998). Sensitivity to change of the roland-morris back pain questionnaire: Part 1. *Physical Therapy, 78*(11), 1186-1196.

Stratford, P. W., Binkley, J. M., & Riddle, D. L. (1996). Health status measures: strategies and analytic methods for assessing change scores. *Physical Therapy, 76*, 1109 – 1123.

Stratford, P. W., Binkley, J. M., & Riddle, D. L. (2000). Development and initial validation of the back pain functional scale. *Spine, 25*(16), 2095-2102.

Stratford, P. W., & Binkley, J. M.. (1997). Measurement properties of the RM-18: a modified version of the roland-morris disability scale. *Spine, 22*(20), 2416-2421.

Streiner, D. L., & Norman, G. (1995). *Measurement Scales. A Practical Guide to their Development and Use* New York: Oxford University Press.

Taylor, S. J. M., Taylor, A. E. P., Foy, M. A. F., & Fogg, A. J. B. F. (1999). Responsiveness of common outcome measures for patients with low back pain. *Spine, 24*(17), 1805.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42.

Terwee, C. B., Roorda, L. D., Dekker, F. W., Bierma-Zeinstra, S. M., Peat, G., Jordan, K., et al. (2010). Mind the MIC; large variation among population and methods. *Journal of Clinical Epidemiology, 63*, 524-534.

Testa, M. A., & Nackley, J. F. (1994). Methods for Quality-Of-Life Studies. *Annual Review of Public Health, 15*(1), 535-559.

Torenbeek, C.E., Caulfield, B., Garrett, R. & van Harten, M. (2001). Current use of outcome measures for stroke and low back pain rehabilitation in five European countries: first results of the ACROSS project. International Journal of Rehabilitation Research, 24 (2), 95 - 101

Turner, J. A. P., Fulton-Kehoe, D. M. P. H., Franklin, G. M. D. M. P. H., Wickizer, T. M. P., & Wu, R. M. (2003). Comparison of the roland-morris disability questionnaire and generic health status measures: a population-based study of workers' compensation back injury claimants. *Spine, 28*(10), 1061-1067.

Tyron, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods* 6 , 371-386

Underwood, M. R., Barnett, A. G., & Vickers, M. R. (1999). Evaluation of two time-specific back pain outcome measures. *Spine, 24*(11), 1104.

van der Roer, N., Ostelo, R. W., Bekkering, G. E., van Tulder, M. W., & De Vet, H. C. (2006). Minimal clinically important change for pain intensity, functional status and general health status in patients with non specific low back pain. *Spine, 31*(5), 578-582.

van Tulder, M., Becker, A., Bekkering, T., Breen, A., Gil del Real, M., Hutchinson, A. (2006). Chapter 3 European guidelines for the management of acute nonspecific low back pain in primary care. *European Spine Journal, 15*(0), s169-s191.

Vickers, A. J., & Altman, D. G. (2001). Statistics Notes: Analysing controlled trials with baseline and follow up measurements. *British Medical Journal, 323*(7321), 1123-1124.

Von Korff, H., Le Resche, T. & Dworkin, R.H. (1993). First onset of common pain symptoms: a prospective study of depression as a risk factor. Pain, *55(*2):251-8.

Waddell, G., Newton, M., Henderson, I., Somerville, D., & Main, C. (1993). A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. . *Pain, 52*(2), 157-168.

Waddell, G., Somerville, D., Henderson, I., & Newton, M. (1992). Objective clinical evaluation of physical impairment in chronic low back pain. *Spine, 17*(6), 617-628.

Walker, B. F. (2000). The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. *Journal of Spinal Disorders & Techniques, 13*(3), 205-217.

Walker, B.F., Muller, R., & Grant, W.D. (2004). Low back pain in australian adults; health provider utilization and care seeking. *Journal of Manipulative and Physiological Therapeutics*, 27, 327-335.

Ward, M. M., Marx, A. S., & Barry, N. (2000). Identification of clinically important changes in health status using receiver operating curves. *Journal of Clinical Epidemiology, 53*, 279-284.

Ware, J.E. & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30,* 473–483

Wells, G., Beaton, D., Shea, B., Boers, M., Simon, L. & Strand, V. (2001). Minimal clinically important differences: review of methods. *The Journal of Rheumatology, 28*(2), 406-412.

Whiting, P., Harbord, R., & Kleijnen, J. (2005). No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Medical Research Methodology, 5*(1), 19.

Whiting, P., Rutjes, A., Reitsma, J., Bossuyt, P., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3*(1), 25.

Whiting, P., Weswood, M. E., Rutjes, A., Reitsma, J., Bossuyt, P. M., & Kleijnen, J. (2006). Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology, 6*(9), 1-8.

Woby, S. R., Urmston, M., & Watson, P. J. (2007). Self-efficacy mediates the relation between pain-related fear and outcome in chronic low back pain patients. *European Journal of Pain, 11*(7), 711-718.

World Health Organization. (2001). International classification of functioning, disability and health (ICF). *Geneva: World Health Organization.*

Wyrwich, K., Metz, S., Kroenke, K., Tierney, W., Babu, A., & Wolinsky, F. (2007). Measuring patient and clinician perspectives to evaluate change in health-related quality of life among patients with chronic obstructive pulmonary disease. *Journal of General Internal Medicine, 22*(2), 161-170.

Wyrwich, K. W., Bullinger, M., Aaronson, N., Hays, R. D., Patrick, D. L., & Symonds, T. (2005). Estimating clinically significant differences in quality of life outcomes. *Quality of Life Research, 14*, 285 - 295.

Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (1999). Further Evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of Clinical Epidemiology, 52*(9), 861-873.

Wyrwich, K. W. M. S., Nienaber, N. A. M. A., Tierney, W. M. M. D., & Wolinsky, F. D. P. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care, 37*(5), 469-478.

APPENDICES

APPENDIX 1

**Health
and
Disability
Ethics
Committees**

22 July 2009

**Ms Roma Forbes
52 Upper Queen St
Newton, Auckland**

Dear Roma

The Lower Limb Tasks Questionnaire; is it useful in the assessment of function in low back pain sufferers?
Investigators: **Roma Forbes, Prof Peter McNair, Peter Larmer.**
Ethics ref: **NTY/09/04/034**
Locations: **Avondale Physiotherapy**

The above study has been given ethical approval by the **Northern Y Regional** Ethics Committee.

Approved Documents
-Questionnaires.
-Participant Information Sheet and Consent Form version 2 dated 15/06/09.
-Physiotherapist Information Sheet version 2 dated 15/06/09.

Accreditation
The Committee involved in the approval of this study is accredited by the Health Research Council and is constituted and operates in accordance with the Operational Standard for Ethics Committees, April 2006.

Final Report
The study is approved until **22 July 2010**. A final report is required at the end of the study and a form to assist with this is available at http://www.ethicscommittees.health.govt.nz. If the study will not be completed as advised, please forward a progress report and an application for extension of ethical approval one month before the above date.

Amendments
It is also a condition of approval that the Committee is advised of any adverse events, if the study does not commence, or the study is altered in any way, including all documentation eg advertisements, letters to prospective participants.

Please quote the above ethics committee reference number in all correspondence.

It should be noted that Ethics Committee approval does not imply any resource commitment or administrative facilitation by any healthcare provider within whose facility the research is to be carried out. Where applicable, authority for this must be obtained separately from the appropriate manager within the organisation.

Yours sincerely

**Amrita Kuruvilla
Northern Y Ethics Committee Administrator**

Email: amrita_kuruvilla@moh.govt.nz

APPENDIX 2

Studies Excluded from Review

Author	Title	Reason for exclusion
Kucukdeveci, Tennant, Elhan & Niyazoglu, 2001	Validation of the Turkish version of the Roland-Morris Disability Questionnaire for use in low back pain.	No calculation of MCID made
Turner, Fulton-Kehoe, Franklin, Wickizer & Wu, 2003	Comparison of the Roland-Morris Disability Questionnaire and Generic Health Status Measures	No calculation of MCID made
Chansirinukor, Maher, Latimer & Hush, 2004	Comparison of the Functional Rating Index and the 18 item Roland Morris Disability Questionnaire; responsiveness and reliability	No calculation of MCID made and RMQ-18 used
Pengel, Refshauge, Maher, 2004	Responsiveness of pain, disability and physical impairment outcomes in patients with low back pain	No calculation of MCID made
Kuijer et al, 2005	Responsiveness of the Roland Morris Disability Questionnaire; consequences of using different external criteria	No calculation of MCID made
Hare-Mortensen, Lauridsen, & Grunnet-Nilsson, 2006	The relative responsiveness of 3 different types of clinical outcome measures on chiropractic patients with LBP.	No calculation of MCID made
Costa, Maher, Latimer, Ferreira, Pozzi & Ribeiro, 2007	Psychometric characteristics of the Brazilian-Portuguese versions of the Functional Rating Index and the Roland Morris Disability Questionnaire.	No calculation of MCID made
Costa et al, 2008	Clinimetric Testing of Three Self-report Outcome Measures for Low Back Pain Patients in Brazil: Which One Is the Best?	No calculation of MCID made
Frost, Lamb & Stewart-Brown, 2008	Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index and the Roland Morris Disability Questionnaire for patients with subacute and chronic LBP	No calculation of MCID made
Ostelo et al, 2008	Interpreting change scores for pain and functional status in low back pain; towards international consensus regarding minimal important change	Review only

Participant Information Sheet

15/06/09

Project: Functional questionnaires; are they useful in the assessment of those with low back pain?

You have been invited to take part in a study.
Please read the information sheet so that you understand what this will involve before making a decision to participate.
You should ask the Investigator any questions necessary to understand the study.
You may have a friend, family or whānau support to help you understand the risks and/or benefits of this study and any other explanation you may require.

Participation in this study is voluntary. If you decide not to participate, this will not affect your ongoing healthcare. If you do decide to participate, you may withdraw at any time without having to give a reason and this will not affect your future care.

What is the purpose of the study?
The aim of this study is to investigate the ability of a questionnaire to accurately measure function in those with low back pain.

Who can take part?

About 50 patients from Avondale Physiotherapy are selected to take part and are chosen by chance, if they are experiencing low back pain. To participate, you need to be over 18 years of age, and give consent to take part.

What happens in the study?

You will be asked to complete two questionnaires on your first visit. This takes 5-10 minutes. After 3-4 weeks of treatment, or when treatment is no longer needed, the two questionnaires are completed again, along with a simple question to check improvement. You do not have to answer all Questions.

What are the potential discomforts and risks?
There is minimal risk associated with this study. Participation in this study will not cost you anything. Taking part in the study will have no effect on the treatment you receive.

What are the benefits?
This study is seeking to find if one questionnaire is as effective as others currently used. Although there is no direct benefit, physiotherapy practice may become more efficient following results of the study.

How is my privacy and confidentiality protected?
Your information is collected and used in accordance with the Privacy Act 1993.

No materials which could personally identify you will be used in any reports on this study.

The following steps are also taken to protect your privacy and confidentiality:

- Questionnaire forms are prepared with unique number only
- Mail-out is packaged and posted
- All names and contact details are deleted from data – leaving only unique number and demographic information

Costs of Participating
Participation in the study is voluntary. There are no costs involved. The expected time to complete the questionnaires is less than 10 minutes.

What happens with the results of the study?
The results of this study may be published in journals and presented at health conferences. It is usual that a delay between the end of the data collection and the presentation of results may occur. The outcomes of this study will be available to you at your request.

Opportunity to consider invitation
You may contact Roma Forbes, romaforbes@hotmail.com phone 09 828 2564 / 09 377 6744, and have your questions answered. You have the right to withdraw from the research at any time, or to ask for your information to be withdrawn.

Participant Concerns
Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor, Dr Peter McNair, peter.mcnair@aut.ac.nz phone 09 921 9999 ext 7146, or Peter Larmer phone 09 921 9999 ext 7322.
If you have any queries or concerns regarding your rights as a participant in this study, you may wish to contact an independent health and disability advocate:
Phone:0800555050
Fax:08002SUPPORT(080027877678)
Email: advocacy@hdc.org.nz'

This study has received ethical approval from the Northern Y Regional Ethics Committee

Ethics Committee, ethics reference number NTY/09/04/034

APPENDIX 4

Consent to Participation in Research

AUT
AUCKLAND UNIVERSITY OF TECHNOLOGY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

15/06/09   Participant Code:

Title of Project: The Lower Limb Tasks Questionnaire; is it useful in the assessment of function in low back pain sufferers?

Project Researchers: Roma Forbes   Position: Primary Researcher

   Dr Peter McNair and Peter Larmer   Position: Primary and Secondary Supervisors

Researcher: Roma Forbes   Student of Master of Health Science of AUT

- I have read and understood the information provided about this research project.

- I have had an opportunity to ask questions and to have them answered.

- I understand that I may withdraw myself or any information that I have provided for this project at any time prior to completion of data collection, without being disadvantaged in any way. If I withdraw, I understand that all relevant information or parts thereof, will be destroyed.

- I agree to take part in this research.

Participant signature:...Date………………

Confidential information:

Participant name:…………………………………………………………….

Contact Address:………………………………………………………….

Contact Phone:…………………………………....

Researcher Contact Details:
Roma Forbes romaforbes@hotmail.com phone 09 828 2564 / 09 377 6744

Project Supervisors Contact Details:
Peter Larmer, peter.larmer@aut.ac.nz phone 09 921 9999 ext 7322
Dr Peter McNair, peter.mcnair@aut.ac.nz phone 09 921 9999 ext 7146

This study has received ethical approval from the Northern Y Regional Ethics Committee Ethics Committee, ethics reference number NTY/09/04/034

AUT
AUCKLAND UNIVERSITY OF TECHNOLOGY
TE WĀNANGA ARONUI O TĀMAKI MAKAU RAU

LOWER LIMB TASKS QUESTIONNAIRE

Patient: _____   **Date:** _____

INSTRUCTIONS

Please rate your ability to do the following activities in the **past 24 hours** by circling the number below the appropriate response.

If you did not have the opportunity to perform an activity in the **past 24 hours**, please make your *best estimate* on which response would be the most accurate.

Please also rate how important each task is to you in your daily life according to the following scale:

1. = Not important
2. = Mildly important
3. = Moderately important
4. = Very important

		NO DIFFICULTY	MILD DIFFICULTY	MODERATE DIFFICULTY	SEVERE DIFFICULTY	UNABLE	IMPORTANCE OF TASK			
1.	Walk for 10 minutes	4	3	2	1	0	1	2	3	4
2.	Walk up or down 10 steps (1 flight)	4	3	2	1	0	1	2	3	4
3.	Stand for 10 minutes	4	3	2	1	0	1	2	3	4
4.	Stand for a typical work day	4	3	2	1	0	1	2	3	4
5.	Get on and off a bus	4	3	2	1	0	1	2	3	4
6.	Get up from a lounge chair	4	3	2	1	0	1	2	3	4
7.	Push or pull a heavy trolley	4	3	2	1	0	1	2	3	4
8.	Get in and out of a car	4	3	2	1	0	1	2	3	4
9.	Get out of bed in the morning	4	3	2	1	0	1	2	3	4
10.	Walk across a slope	4	3	2	1	0	1	2	3	4

		NO DIFFICULTY	MILD DIFFICULTY	MODERATE DIFFICULTY	SEVERE DIFFICULTY	UNABLE	IMPORTANCE OF TASK			
1.	Jog of 10 minutes	4	3	2	1	0	1	2	3	4
2.	Pivot or twist quickly while walking	4	3	2	1	0	1	2	3	4
3.	Jump for distance	4	3	2	1	0	1	2	3	4
4.	Run fast/sprint	4	3	2	1	0	1	2	3	4
5.	Stop and start moving quickly	4	3	2	1	0	1	2	3	4
6.	Jump upwards and land	4	3	2	1	0	1	2	3	4
7.	Kick a ball hard	4	3	2	1	0	1	2	3	4
8.	Pivot or twist quickly while running	4	3	2	1	0	1	2	3	4
9.	Kneel on both knees for 5 minutes	4	3	2	1	0	1	2	3	4
10.	Squat to the ground/floor	4	3	2	1	0	1	2	3	4

The Roland-Morris Disability Questionnaire

Name: Date:

When your back hurts, you may find it difficult to do some of the things you normally do.
Mark only the sentences that describe you today.

- ☐ I stay at home most of the time because of my back.
- ☐ I change position frequently to try to get my back comfortable.
- ☐ I walk more slowly than usual because of my back.
- ☐ Because of my back, I am not doing any jobs that I usually do around the house.
- ☐ Because of my back, I use a handrail to get upstairs.
- ☐ Because of my back, I lie down to rest more often.
- ☐ Because of my back, I have to hold on to something to get out of an easy chair.
- ☐ Because of my back, I try to get other people to do things for me.
- ☐ I get dressed more slowly than usual because of my back.
- ☐ I only stand up for short periods of time because of my back.
- ☐ Because of my back, I try not to bend or kneel down.
- ☐ I find it difficult to get out of a chair because of my back.
- ☐ My back is painful almost all of the time.
- ☐ I find it difficult to turn over in bed because of my back.
- ☐ My appetite is not very good because of my back.
- ☐ I have trouble putting on my sock (or stockings) because of the pain in my back.
- ☐ I can only walk short distances because of my back pain.
- ☐ I sleep less well because of my back.
- ☐ Because of my back pain, I get dressed with the help of someone else.
- ☐ I sit down for most of the day because of my back.
- ☐ I avoid heavy jobs around the house because of my back.
- ☐ Because of back pain, I am more irritable and bad tempered with people than usual.
- ☐ Because of my back, I go upstairs more slowly than usual.
- ☐ I stay in bed most of the time because of my back.

HOW WOULD YOU RATE YOUR CHANGE SINCE YOUR FIRST VISIT?

Please indicate on the following scale:

Completely Recovered

Much Improved

Improved

Slightly Improved

No Change

Slightly Worse

Much Worse