# A SYSTEMATIC REVIEW OF HEAD-TO-HEAD COMPARISON STUDIES OF THE ROLAND-MORRIS AND OSWESTRY MEASURES' ABILITIES TO ASSESS CHANGE

# A SYSTEMATIC REVIEW OF HEAD-TO-HEAD COMPARISON STUDIES OF THE ROLAND-MORRIS AND OSWESTRY MEASURES' ABILITIES TO ASSESS CHANGE

by

ANASTASIA N. L. NEWMAN

H.B. Kin, MSc (PT)

A Thesis
Submitted to the School of Graduate Studies
In Partial Fulfillment of the Requirements
For the Degree
Master of Science
Rehabilitation Science

McMaster University
Hamilton, Ontario

| | |
|---|---|
| MASTER OF SCIENCE (2011) | MCMASTER UNIVERSITY |
| Department of Rehabilitation Science | Hamilton, Ontario, Canada |

TITLE: A Systematic Review of Head-to-Head Comparison Studies of the Roland Morris and Oswestry Measures' Abilities to Assess Change

AUTHOR: Anastasia N.L. Newman H.B. Kin., MSc (PT)

SUPERVISOR: Professor Paul W. Stratford PT, MSc

NUMBER OF PAGES: 88

# A SYSTEMATIC REVIEW OF HEAD-TO-HEAD COMPARISON STUDIES OF THE ROLAND-MORRIS AND OSWESTRY MEASURES' ABILITIES TO ASSESS CHANGE

## Abstract

Low back pain (LBP) is a common musculoskeletal condition that can lead to pain, functional limitations and disability. Due to the prevalence of LBP, multiple self-reported outcome measures have been developed, which have resulted in redundancy in the literature. Two frequently used outcome measures are the Roland Morris Questionnaire (RMQ) and the Oswestry Disability Index (ODI). Few authors have performed head-to-head comparison studies to determine which of these outcome measures are the most successful at measuring sensitivity to change. The purpose of this thesis was to answer the question: Is there a difference in the sensitivity to change between the RMQ and the ODI in their ability to measure pain-related functional status in persons with low back pain?

The first part of this thesis involves a systematic review of head-to-head comparison studies to determine the difference in the sensitivity to change of the RMQ and the ODI. Five databases were searched and nine articles were located. The second part of this thesis entails the development of a quality criteria form to evaluate head-to-head comparison studies, which was applied to the nine eligible articles.

The third aspect of this research was to perform a head-to-head comparison study of the RMQ and the ODI using data from the nine head-to-head studies. A small but significant difference was noted in favour of the RMQ in terms of the Spearman rank correlation coefficient between its change scores and the reference standard (Z = 2.36, p = 0.018; Z = 3.28, p = 0.001) and also in the Receiver Operating Characteristic curve area ($X^2_1$ = 8.58, p = 0.003).

The results of this systematic review and head-to-head comparison study found a small but significant difference in favour of the RMQ in its ability to measure the sensitivity to change in patients with LBP.

## Dedication

*to my husband Adam*
*for his unending love, support, encouragement*
*and technical wizardry*

## Acknowledgements

Professor Paul W. Stratford
*for his continued support, feedback, encouragement, friendship and calming reassurance*

Dr. Lori Letts
*for her guidance, patience and wealth of knowledge*

Gregory Spadoni
*for his insights, clinical knowledge and perspectives and sense of humour*

Deborah Kennedy
*for her thoughtful comments and encouragement during the defense*

Dr. Anne Mannion, Dr. Sandra Beurskens, Prof. Megan Davidson and Prof. Margreth Grotle
*for their willingness to share their knowledge and information*
*and for supporting Physiotherapy research*

**Contents**

## List of Tables

## List of Figures

# Chapter 1: Introduction and Background

## 1.1 Introduction to Chapter

Low back pain (LBP) is a persistent cause of morbidity and disability. Due to the frequency of LBP, there has been a proliferation of self-reported outcome measures used to evaluate a patient's status. This chapter reviews the epidemiology and social impact of low back pain, defines an outcome measure, and describes the related psychometric properties that are integral to conducting outcome measures research.

## 1.2 Epidemiology and Social Impact of Low Back Pain

Low back pain is one of the most common and costly health conditions (Cleland, Gillani, Bienen, & Sadosky, 2011). Recent statistics from the United States indicate that approximately 50,000,000 American adults have experienced low back pain "in the past three months" (Cleland et al., 2011, p. 1). It is estimated that 80 to 85 percent of individuals will experience LBP at some point in their lifetime (Deyo & Centor, 1986; Hoy, Brooks, Blyth, & Buchbinder, 2010) and that approximately two to five percent of people will develop chronic LBP lasting greater than three months (Hoy et al., 2010). The impact of LBP is multi-factorial and is a major cause of morbidity, including pain, activity and participation limitations, career burden, use of health care resources, and financial stress (Hoy et al., 2010).

## 1.3 Introduction to Outcome Measures and Relevant Terminology

An outcome measure is defined as a measurement tool used to document change in one or more constructs over time (Finch, Brooks, Stratford, & Mayo, 2002). Two essential properties of outcome measures are their ability to discriminate among persons at a single point in time and their ability to assess change over time (Finch et al., 2002; Kirshner & Guyatt, 1985). For an outcome measure to be useful in a specific clinical population its results must demonstrate acceptable levels of reliability, validity, sensitivity to change, and responsiveness (Finch et al., 2002).

### 1.3.1 Reliability

Reliability is defined as the "degree to which test scores are free from errors of measurement" (Domholdt, 2000, p. 255) and is a necessary, but not a sufficient, condition for a measure to be considered valid (Domholdt, 2000). Two specific requirements must be met for an outcome measure to be deemed reliable (Finch et al., 2002). First, a reliable measure must provide consistent results with small errors of measurement (Finch et al., 2002). Second, it must be capable of differentiating between subjects with whom the measure is being used (Finch et al., 2002).

Reliability can be expressed in two different ways: relative reliability and absolute reliability. Relative reliability examines the relationship between two or more sets of repeated measures (Finch et al., 2002) and is based on the presumption that if a measure is truly reliable, individual measurements within a group will maintain their position in that group on repeated measurements (Domholdt, 2000). Relative reliability, which is measured using an intraclass correlation coefficient, indicates the strength of an association between repeated measurements of the outcome measure or variable of interest (Domholdt, 2000).

Absolute reliability expresses the extent to which a score will vary on repeated measurements (Domholdt, 2000) and expresses the measurement error in the same units as the original measurement (Finch et al., 2002). Absolute reliability is quantified using the standard error of the measurement (SEM) (Domholdt, 2000).

There are several types of reliability that are frequently assessed and evaluated in quantitative research. The four most common forms of reliability are test-retest reliability, internal consistency, interrater reliability and intrarater reliability. Test-retest reliability involves having participants take the same test on two or more occasions (Domholdt, 2000). Internal consistency is based on parallel assessments of subjects at an instant in time and is typically used when multi-item questionnaires or performance measure scores are summarized into a single score (Finch et al., 2002). Interrater reliability evaluates the consistency of performance between different raters in assigning scores to the same subject sample (Domholdt, 2000). Intrarater reliability is the consistency with which one rater assigns a score to the same subject sample on two separate occasions.

### 1.3.2 Validity

The interpretation of the term validity has witnessed several evolutions since the 1930s, and has continued to evolve over the last several decades. In the formative years of validity theory, validity was described by Garrett (1937) as “the fidelity with which it measures what it purports to measure” (Shultz, Riggs, & Kottke, 1998, p. 2). In the late 1940s and early 1950s, validity was believed to be mostly criterion-based and relied on a single correlation coefficient between the test score and some external criterion to establish a test's validity (Shultz et al., 1998). By the mid-1950s, several types of validity were being described in the literature, including intrinsic validity, factorial validity and empirical validity (Shultz et al., 1998).

In an effort to create consistency in the literature, the 1954 Technical Recommendations for Psychological Tests and Diagnostic Techniques proposed four categories of validity: predictive, concurrent, content and construct validity (Landy, 1986). Eventually, predictive and concurrent validity were combined under the term criterion-related validity (Landy, 1986). These three types of validity, criterion, content and construct, became known as the Trinitarian view (Landy, 1986; Shultz et al., 1998). Since much of the work done to create the Trinitarian view was done in the field of Psychology, the definitions of the three types were based on psychological tests and situations. These three types of validity were seen as independent of each other and therefore needed to be assessed on an individual basis (Norman & Streiner, 2008). Furthermore, because construct validity could be analyzed in a multitude of ways, it was further divided into several subcategories, including convergent validity, discriminant validity and trait validity (Norman & Streiner, 2008). The emergence of these classifications of validity led to debate in the research community as to the specific type of validity being assessed or measured in research (Norman & Streiner, 2008).

Two important trends emerged that would help to yet again modify the views on validity. First, a major conceptual shift occurred in regards to the object of validation (Norman & Streiner, 2008). It was initially believed up to the 1960s that it was the scale or its measurement properties that was found to be "valid" (Norman & Streiner, 2008). In 1971, Cronbach suggested that it's not the measure itself that is valid, but the characteristics of the subjects being assessed and the scores they obtained on the specific measure (Cronbach, 1971).

As Landy (1986) stated: “Validation processes are not so much directed toward the integrity of tests as they are directed toward the inferences that can be made about the attributes of people who have produced those test scores” (Landy, 1986, p. 1186). This means, in essence, that we cannot state that the scale is valid, only the extent to which the scale has been shown to be valid with a specific population and in a specific context (Norman & Streiner, 2008).

Second, the Trinitarian view of validity was questioned for its oversimplification and supposed convenience (Shepard, 1993). By 1974, the American Psychological Association (APA) recognized the interrelatedness of the three different classifications of validity. At the 1974 APA National Council on Measurement in Education, it was recognized that “these aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation” (Shepard, 1993, p. 12). This realization led to the development of the Unitarian view: that the validation of a measure is simply hypothesis testing (Landy, 1986). Messick, one of the proponents for this change in conceptualization, stated in 1988 that all measurement “should be construct-referenced because construct interpretation undergirds all score-based inferences – not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores” (Messick, 1988, p. 35).

This new vision of validity has been accepted by the Standards for Educational and Psychological Testing in 1999, solidifying the acceptance of validity as hypothesis testing. However, despite this change in terminology and generalized acceptance of this paradigm shift, textbooks and articles still refer to the many types of validity instead of the overarching term of construct validity (Norman & Streiner, 2008). Eventually these subcategories of validity will be eliminated and validity will be recognized as hypothesis or construct testing.

### 1.3.3 Sensitivity to Change and Responsiveness

A challenge facing those interested in the outcome measurement literature is the lack of agreement concerning terminology. Specifically, considerable controversy exists centering on the terms sensitivity to change and responsiveness (Guyatt, Deyo, Charlson, Levine, & Mitchell, 1989; Hays & Hadorn, 1992; Liang, 2000). Guyatt et al (1989) state that sensitivity to change and responsive are synonymous and are both defined as the ability of a measure to detect clinically important changes over time (Guyatt et al., 1989). In contrast, Liang (2000) makes a distinction between responsiveness and sensitivity to change by defining the latter as the ability of an instrument to measure change in a state regardless of whether it is relevant or meaningful to the decision maker while responsiveness is the ability of a tool to measure meaningful or clinically important change in a clinical state (Liang, 2000). Liang goes on to state that "sensitivity to change is a necessary but insufficient condition for responsiveness" (p. 85). Yet other authorities convincingly argue that sensitivity to change and responsiveness are components of validity (Hays & Hadorn, 1992). Their position is that if the goal of a measurement tool is to detect important change, then the extent to which it successfully accomplishes this goal comments on the measure's validity. For the purpose of this thesis, the term sensitivity to change, as defined by Liang (2000), will be used.

## 1.4 Outcome Measures and Low Back Pain

Due to the high prevalence of LBP and its financial and social consequences, the development and implementation of outcome measures designed to assess and quantify a patient's functional limitations has resulted in a proliferation of similar and competing measures in the literature over the last several decades. In 2005, Grotle and colleagues performed a systematic review and identified twenty-eight different outcome measures and eight subsequent revised measures for evaluating LBP-specific function and disability (Grotle, Brox, & Vollestad, 2005). This redundancy in outcome measures has led to confusion in the clinical setting as to which measurement tool is the most appropriate to use in the diverse LBP population (Cleland et al., 2011). One of the reasons for the many competing measures is that as new measures are developed, older and less proficient tools are not eliminated from practice.

The plethora of outcome measures has also led to issues in the research community around standardization of results when attempting to compare different measures between studies (Frost, Lamb, & Stewart-Brown, 2008; Grotle, Brox, & Vollestad, 2004). In an attempt to assist with the comparison of results between trials and to allow for the amalgamation of data from various studies, a group of researchers suggested that outcome measures for LBP contain five specific core domains: pain, back specific function, work disability, generic health status and patient satisfaction (Bombardier, 2000; Deyo et al., 1998). This same group also recommended a core set of outcome measures that are suggested for use in LBP research and practice. Two of these recommended measures include the Roland Morris Questionnaire (RMQ) and the Oswestry Disability Index (ODI) (Frost et al., 2008). Although the RMQ and the ODI have been examined and tested comprehensively in terms of reliability, validity, sensitivity to change and responsiveness, there is a lack of agreement in the literature as to whether one measure is superior to the other. Applying the framework of a systematic review, this thesis compares the sensitivity to change of the RMQ and the ODI and attempts to determine if there is a difference in the abilities of these measures to detect change.

## 1.5 Chapter Summary

This chapter provided a brief epidemiological review and background of LBP and an introduction into the key terminology in outcomes research. Important terms, such as reliability, validity, sensitivity to change and responsiveness were defined. As there are numerous outcome measures in print that have been created, validated and are in use for patients with LBP, there is a need to evaluate these measures in relation to each other and to determine which ones are the most sensitive to change and clinically important.

# Chapter 2: Introduction to the Oswestry Disability Index and the Roland Morris Questionnaire

## 2.1 Introduction to Chapter

There are more than 28 outcome measures in publication for assessing patients' with low back pain. The Oswestry Disability Index (ODI) and the Roland Morris Questionnaire (RMQ) are two of the most frequently used self-report measures in both clinical and research settings (Roland & Fairbank, 2000). This chapter will first address the oversaturation of outcome measures in the literature and then outline the history, development and properties of both the ODI and the RMQ.

## 2.2 Low Back Pain Outcome Measure Oversaturation

With more than two dozen outcome measures currently in circulation for the assessment of low back pain it can be difficult for researchers and clinicians to decide which one to use. This redundancy in the literature results in difficulty interpreting the various measures in terms of their scores and their clinical relevance and limits comparison between outcome measurement tools and the interventions being evaluated (Deyo et al., 1998). In 1997, a task force of multinational researchers met in the Netherlands to explore whether a proposed set of standardized outcome measures for low back pain could be considered (Deyo et al., 1998). These researchers felt that a core set of outcome measures would allow for easier comparison between studies, allow for pooling of data from multiple studies to determine a treatment effect and encourage the cooperation between multiple sites (Deyo et al., 1998).

The two outcome measures recommended by the 1997 task force for assessing back-related function were the ODI and the RMQ (Deyo et al., 1998). The ODI and the RMQ are the most commonly used measurement tools in both the research and the clinical fields relating to low back pain and dysfunction (Davies & Nitz, 2009). Both the ODI and the RMQ can be used to evaluate a patient's pain-related functional status and to monitor the effects of physiotherapy interventions. Interestingly, despite the original authors of the ODI and the RMQ stating that these measures assess pain-related functional status, the 1997 Netherlands task force separated pain and function in their core set. The ODI and the RMQ are recommended for assessing functional status only (Deyo et al., 1998). However, both measures include questions that assess how back pain affects function, and this recommendation to separate pain from function seems to contradict the initial purpose for the development of the ODI and the RMQ.

Because the ODI and the RMQ have been thoroughly investigated for reliability, validity and responsiveness, this thesis includes a comparison of their sensitivity to change by presenting a systematic review to answer the research question: Is there a difference in the sensitivity to change between the RMQ and the ODI in their ability to measure pain-related functional status in persons with low back pain? A quality criteria form was created and used to analyze head-to-head comparison studies of the RMQ and the ODI.

## 2.3 Oswestry Disability Index

The Oswestry Disability Index (ODI), a self-report outcome measurement tool, was originally developed over a four-year period in the late 1970s by a group of physicians and allied health professionals as a condition-specific measure to assess low back pain related symptoms (Fairbank, Davies, Couper, & O'Brien, 1980; Fairbank & Pynsent, 2000). Patients with low back pain were interviewed by an orthopedic surgeon, an occupational therapist and a physiotherapist and asked to identify tasks and aspects of daily life that were disrupted by low back pain (Roland & Fairbank, 2000). Ten items were selected for the first version of the ODI, version 1.0, which were believed to be the most pertinent to patients with low back pain (Fairbank et al., 1980). These items included pain intensity, personal care, lifting, walking, sitting, standing, sleeping, sex life, social life and travelling (Fairbank et al., 1980). The instructions then asked a client to "mark the box which most closely describes your problem" (Fairbank et al., 1980, p. 272). In the late 1980s, the Medical Research Council (MRC) group created version 2.0 of the ODI by changing the instructions to "mark one box only in each section that most closely describes you today", adding in "today" (Roland & Fairbank, 2000). The original ODI did not specify any timeframe for the symptoms. Version 2.0 of the ODI is generally accepted by the original authors as the preferred version to use in a clinical setting (Roland & Fairbank, 2000).

Each section of the ODI contains six statements that each represents increasing levels of disability. Scores for each statement range from zero to five, with zero representing minimal disability and five representing significant disability (Fairbank & Pynsent, 2000). The maximum score attainable is 50 points. Patients are asked to select only one statement in each section. If more than one statement is chosen, the highest scoring statement is taken as the true response (Fairbank et al., 1980). Once completed, the scores for the 10 sections are added together and converted into a percentage, with the totaled points divided by 50 and then multiplied by 100 (Fairbank & Pynsent, 2000). Should any item be left blank, the score is simply amended by adjusting the denominator (Fairbank et al., 1980). The original article provides a breakdown of the percentage scores: zero to 20 percent (minimal disability); 20 to 40 percent (moderate disability); 40 to 60 percent (severe disability); 60 to 80 percent (crippled); 80 to 100 percent (bed-bound or exaggerated symptoms) (Fairbank et al., 1980).

Version 1.0 of the ODI was initially investigated in a 25 subject study in which patients with their first episode of low back pain and who were expected to recover spontaneously were asked to complete the ODI at weekly intervals over a three week period (Fairbank et al., 1980). The scores on the ODI version 1.0 reflected the overall improvement in symptoms felt by the subjects and this reached statistical significance ($p < 0.005$). In a separate study, 22 patients with a history of chronic low back pain were asked to complete the ODI version 1.0 on two separate occasions, one day apart, under similar conditions. The test-retest reliability was found to be $r = 0.99$ ($p < 0.001$) between the two tests (Fairbank et al., 1980).

As version 2.0 is the preferred edition of the ODI, even by the original authors, this thesis focuses on this modified version. The ODI v.2.0 has been frequently investigated to determine its psychometric properties. Internal consistency was found to range from $r = 0.76$ to $r = 0.87$, while test-retest reliability was found to be $r = 0.91$ over four days and $r = 0.83$ over seven days (Fairbank & Pynsent, 2000). It has also been found to correlate strongly with the Quebec Back Scale ($r = 0.77$) (Kopec et al., 1995). The ODI v.2.0 can be found in Appendix 1.

## 2.4 Roland Morris Questionnaire

The Roland Morris Questionnaire (RMQ), a self-report outcome measurement tool, was initially developed by Martin Roland and Richard Morris in 1983 as part of a research study investigating the natural history of low back pain. The RMQ was created by selecting twenty four statements from the Sickness Impact Profile (SIP), a one-hundred and thirty-six item health status measure that assesses multiple aspects of both physical and mental function (Roland & Fairbank, 2000). Sickness Impact Profile statements were selected for the RMQ if it was believed that they specifically related to aspects of physical function that may be influenced by low back pain (Roland & Fairbank, 2000). The phrase “because of my back pain” was added to the beginning or the end of each statement (Roland & Morris, 1983; Roland & Fairbank, 2000). A six-point pain scale in the form of a thermometer accompanied the outcome measure, with possible answers ranging from “no pain at all” at the bottom of the thermometer to “the pain is almost unbearable” at the top (Roland & Morris, 1983). This pain scale is no longer used and the original authors now recommend the use of the SF-36 pain scale instead (Roland & Fairbank, 2000). The RMQ is noted to be short, simple to complete, and easily understood by patients (Roland & Fairbank, 2000). The measure can be completed and scored in approximately five minutes (Roland & Morris, 1983). Scoring of the RMQ consists of giving one point for each statement that was checked by the patient and zero points for statements left blank. Scores range from zero (no disability) to twenty-four (maximum disability) (Roland & Fairbank, 2000). Stratford and colleagues (1994) noted that the RMQ had fewer incomplete or ambiguous responses than the Oswestry Disability Index (Stratford, Binkley, Solomon, Gill, & Finch, 1994). The RMQ can be found in Appendix 2.

Preliminary investigations into the psychometric properties of the RMQ found it to be a reliable and valid measurement tool (Roland & Morris, 1983). Test-retest reliability was quoted at r = 0.91 for scoring between two occasions. The RMQ was shown to have good agreement with the six-point pain scale as well as moderate agreement with physical signs of functional limitations by physicians (Roland & Morris, 1983). More recent reliability values have been estimated as follows: Internal Consistency ranging from r = 0.84 to r = 0.93 (Roland 2000); Test-

Retest reliability ranges from r = 0.91 (same day) (Roland & Morris, 1983), r = 0.88 (one week) (Johansson & Lindberg, 1998) and r = 0.83 (three weeks) (Deyo & Centor, 1986). To assess construct validity, the RMQ has been compared to other health status and low back pain measures and has correlated well with the Quebec Back Scale (r = 0.77) (Kopec et al., 1995), the Sickness Impact Profile (r = 0.82 – 0.91) (Jensen, Strom, Turner, & Romano, 1992) and the ODI (r = 0.77) (Fairbank & Pynsent, 2000).

## 2.5 Relationship between the Oswestry Disability Index and the Roland Morris Questionnaire

Both the ODI and the RMQ are widely used outcome measures that assess pain-related functional status. Roland and Fairbank (2000) state that both outcome measures have been extensively tested and both are applicable for use in a variety of clinical and research settings (Roland & Fairbank, 2000). Overall, the ODI and the RMQ are relatively similar and the differences between these two measures are relatively small (Roland & Fairbank, 2000).

The International Classification of Functioning, Disability and Health (ICF), is both a framework and classification system of health and health-related domains (World Health Organization, 2011). It provides a unified and standard language that can be used to describe health and health-related conditions. The ICF domains are classified using two lists: a list of body functions and structures and a list of activity and participation (World Health Organization, 2011). Environmental factors are also considered since an individual's functioning and disability occurs in a context (World Health Organization, 2011). Developed by the World Health Organization (WHO) and endorsed by all 191 WHO members in 2001 as the international standard for describing health and disability, the ICF does not classify people but describes the health of each individual according to how they are functioning within their environment (World Health Organization, 2011).

Both the ODI and the RMQ incorporate multiple aspects of each of the four domains of the ICF.  In 2004, Cieza et al reported the results of a consensus process which was performed to define the typical spectrum of problems in patients with low back pain (Cieza et al., 2004).  This study determined seventy-eight categories of the ICF domains specific to low back pain (Cieza et al., 2004).  The following two tables show which of the categories from each of the four domains the ODI and the RMQ touch on.  Neither outcome measure comprehensively assesses both body structures and environmental factors.

**Table 2-1: ICF Categories Included in the Oswestry Disability Index and the Roland Morris Questionnaire**

| | ODI | RMQ |
|---|---|---|
| **Body Functions** | Sensation of pain<br>Exercise tolerance functions<br>Muscle endurance functions<br>Sleep functions | Sensation of pain<br>Emotional functions<br>Muscle power functions<br>Sleep functions |
| **Body Structures** | None | None |
| **Activities and Participation** | Maintaining a body position<br>Lifting and carrying objects<br>Walking<br>Dressing<br>Family relationships<br>Intimate relationships | Lifting and carrying objects<br>Changing basic body positions<br>Walking<br>Doing housework<br>Dressing |
| **Environmental Factors** | Travel functions | Stair use |

In terms of floor effects (a measurement phenomenon in which an instrument cannot register greater declines in scores for the participants of interest)(Domholdt, 2000)  and ceiling effects (a measurement phenomenon in which an instrument cannot register gains in scores for the participants of interest) (Domholdt, 2000) the RMQ tends to have a higher proportion of patients scoring in the top half of the scale as compared to the ODI, mostly due to the fact that the RMQ is scored out of twenty-four point and the ODI out of fifty points (Roland & Fairbank, 2000). However, the RMQ tends to discriminate between patients scoring at the lower end of the scale as compared to the ODI (Roland & Fairbank, 2000). Both authors of the original studies recommend the use of the ODI for patients who are likely to have sustained severe disability, while the RMQ is recommended for patients who are expected to have relatively little disability (Roland & Fairbank, 2000).

The ODI and the RMQ have been tested comprehensively since their developments in the 1980's. Scores from both measures have been found to be highly correlated with similar test-retest reliability and internal consistency (Kopec et al., 1995). Maughan and Lewis (2010) compared the ODI and the RMQ using an ROC curve analysis and found the area under the curve (AUC) to be 0.64 and 0.67 respectively (Maughan & Lewis, 2010), while Beurskens et al (1996) found the AUC for the ODI and the RMQ to be 0.76 and 0.93 respectively (Beurskens, de Vet, & Koke, 1996). Frost et al (2008) found the AUC for the ODI to be 0.75 while the RMQ value was 0.69 (Frost et al., 2008)(Frost et al., 2008). In a 2004 article by Grotle et al, the RMQ had a greater AUC in both an acute and a chronic low back pain groups (0.93 and 0.83 respectively) as compared to the ODI (0.87 and 0.75 respectively) (Grotle et al., 2004). Overall, the RMQ tends to have slightly higher AUC than the ODI.

Effect size (ES) values have been quoted by several authors. Beurskens et al (1996) found the ES of the RMQ to be 2.02 while the ES of the ODI was noted to be 0.80 (Beurskens et al., 1996). In a 2008 prospective cohort study by Frost et al, a large ES was noted for the ODI (-0.88 to -1.00) although only a moderate ES of -0.70 to -0.74 was found for the RMQ(Frost et al., 2008).

In patients experiencing acute low back pain reporting an improvement in status, Grotle et al (2004) found a standardized response mean (SRM) of 1.5 for the RMQ and an SRM of 1.4 for the ODI (Grotle et al., 2004). In acute patients who reported no change in status, the SRM for the RMQ was 0.3 and the SRM for the ODI was 0.7 (Grotle et al., 2004). In a chronic population, the SRM for the change group was 1.1 and 1.0 for the RMQ and the ODI respectively while the unchanged group was -0.1 and 0.2 respectively for the RMQ and the ODI (Grotle et al., 2004). The standard error of measurement (SEM) is often believed to be the best method for calculating important change (Maughan & Lewis, 2010). Maughan et al (2010) found the SEM for the RMQ to be 1.78 and the SEM for the ODI to be 6.06 (Maughan & Lewis, 2010).

Responsiveness estimates for the RMQ and the ODI have also been routinely investigated. Ostelo et al (2008) proposed a minimally important clinical difference (MCID) of 5 points for the RMQ and 10 points for the ODI (Ostelo, Deyo, Stratford, Waddell, Croft, Von Korff, Bouter, & de Vet, 2008b), while Maughan et al (2010) found an MCID of 3.5 points for the RMQ and 7.5 points for the ODI based on an ROC analysis (Maughan & Lewis, 2010).

## 2.6 Chapter Summary

The ODI and the RMQ are two well-established and highly-used self-report outcome measures for low back pain in research and clinical settings. They have been routinely tested for both reliability and validity and are two of the most frequently used outcome measures in low back pain research. It is because of their frequency in the literature that these two outcome measures were chosen for the head-to-head comparison of this thesis.

# Chapter 3: Review of Methodologies Applied in Head-to-Head Comparison Studies of Competing Outcome Measures

## 3.1 Chapter Introduction

With the increasing focus on the development and implementation of health status outcome measures in the last several decades, there has often been uncertainty in the literature as to the most appropriate study design and analysis techniques to assess a measure's ability to detect change (Stratford & Riddle, 2005). The result of this confusion is the use of multiple analysis techniques, as opposed to the selection of the most appropriate statistical test for the data. It is likely that this occurs because for many important outcomes (pain, functional status, and health related quality of life) there exists no gold standard for assessing change (Stratford & Riddle, 2005). This chapter provides a brief review of typical study designs and analysis techniques used to evaluate outcome measures' abilities to assess change over time.

## 3.2 Typical Study Designs

The typical study designs used to evaluate outcome measures can be divided into single-group and multiple-group designs (Stratford, Binkley, & Riddle, 1996), each of which has several subdivisions.

### 3.2.1 Single-Group Designs

The most basic single-group design is the before and after design, as seen in figure 3-1. Patients are expected to undergo a change in health status from the first to the second assessment. The time interval between the two assessments and any intervention applied during this period serve as the constructs for change (Stratford et al., 1996). It is the ability of the specific outcome measure to detect a difference between the two assessments that represents its sensitivity to change or responsiveness.

There are two limitations with this type in single-group design. First, if no change between assessment scores is identified, it is unknown whether this is due to the measure's inability to detect change or if the time interval between measurements or the intervention used was at fault, i.e. that no change actually occurred (Stratford et al., 1996). Second, this type of design does not allow for the evaluation of patients who remain stable during the time frame. Stratford et al (1996) believe this to be the weakest study design (Stratford et al., 1996).

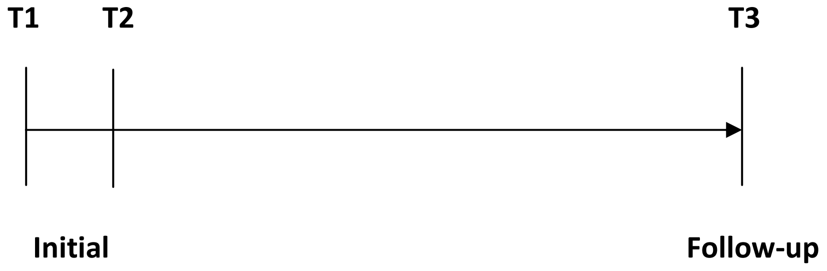
**Figure 3-1: Single Group Design Type 1**

T1
T2
Initial
Follow-up

The second type of single-group design involves taking patient measurements at three points in time, which is proposed to remediate the inability of the first design type's ability to measure change and stability in patient status (Stratford et al., 1996). The interval between time 1 and time 2 is less than that of the interval between time 2 and time 3 (Figure 3-2). It is hypothesized that the amount of patient change between time 1 and time 2 will be less than the amount of change between time 2 and time 3. The stability of a patient's status is theorized to be assessed during the first and second assessments while true patient change is measured during the second and third assessments.

The main limitation of this study design is that patient stability is measured over a shorter time period than patient change which may underestimate the magnitude of random variability in patient status that can occur over a longer period of time in patients whose health status does remain relatively stable (Stratford et al., 1996). This study design assumes that there is patient stability between time 1 and time 2 and patient status change between time 2 and time 3.

**Figure 3-2: Single Group Design Type 2**

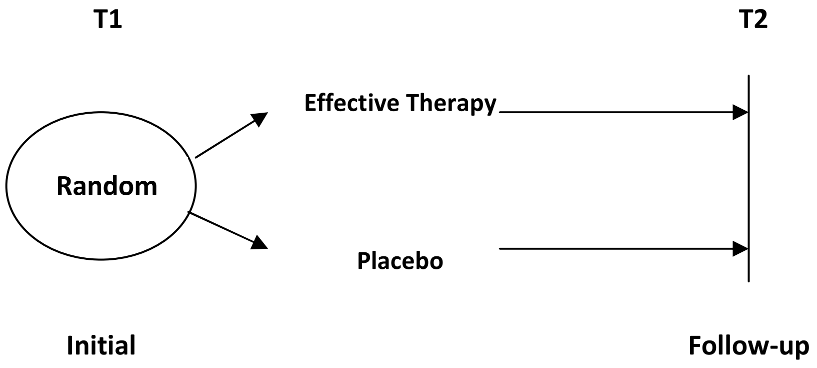


**3.2.2 Multiple-Group Designs**

Multiple-group design improves on the single-group design by providing a more accurate assessment of change and stability over the same amount of time (Stratford et al., 1996). There are three subtypes of the multiple-group designs that are linked by two common factors. First, it is expected that the health status of the patients will change by differing amounts. Second, it is important to ensure that each group of patients is roughly equal in size to ensure statistical efficiency. It is the outcome measure’s ability to detect differing amounts of change between the groups that is being assessed in the multiple-group designs.

The first type of multiple-group design requires that a previously proven and effective treatment exists (Stratford et al., 1996) as patients are randomly assigned to either the intervention or placebo/control group. Each patient is initially assessed at time one and then randomly assigned to either group. Between time one and time two the intervention or placebo treatment is provided to the patients. All patients are assessed at time two at the end of the study. The theory behind this design is that those patients who experienced the proven intervention will demonstrate greater improvements in their status as well as score higher on the outcome measure than those patients who received the placebo. This design is depicted in figure 3-3.

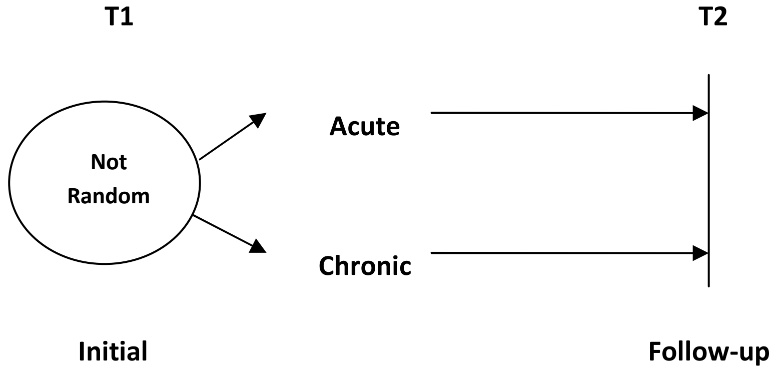
**Figure 3-3: Multiple Group Design Type 1**



The two limitations with this type of multiple-group design are the need for having an effective treatment to offer patients and the ethical concerns with denying the placebo group a known effective intervention (Stratford et al., 1996).

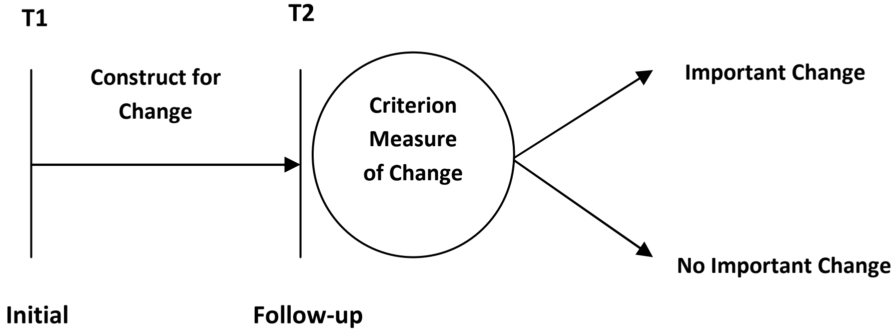
The second type of multiple-group design includes known subgroups or categories of the condition being studied. Patients are divided into two groups based on the known severity or acuity of the condition. For example, patients with low back pain could be divided into an acute and a chronic group (Stratford et al., 1996). Patients are once again assessed at two time points, as shown in figure 3-4. The extent to which the outcome measure in question can discriminate between the two patient subgroups indicates its ability to detect change.

Figure 3-4: Multiple Group Design Type 2



The final type of multiple-group design utilizes the addition of a reference standard of change to help assess a change in a subject's health status. Subjects are assessed at two time points, and the reference standard is given to each subject after the final assessment to classify the extent of change that is believed to have truly occurred. The outcome measure under investigation is also applied after the final assessment and the more proficient it is, the better its ability to correlate highly with the reference standard (Stratford et al., 1996). Figure 3-5 depicts this study design type.

Figure 3-5: Multiple Group Design Type 3



## 3.3 Typical Analytic Techniques

Prior to discussing the typical statistical techniques for sensitivity to change, it is necessary to distinguish the difference between distribution-based and anchor-based methods of analysis. Distribution-based methods express the observed change in a standardized metric (Ostelo, Deyo, Stratford, Waddell, Croft, Von Korff, Bouter, & de Vet, 2008a) and measure the statistical significance of the change scores (Maughan & Lewis, 2010). Examples of distribution-based methods include the standard error of measurement (SEM), effect size (ES) and standardized response means (SRM) (Grotle et al., 2004). The major disadvantage of distribution-based analysis is that the results are purely statistical in nature and do not necessarily reflect clinically important changes (Ostelo, Deyo, Stratford, Waddell, Croft, Von Korff, Bouter, & de Vet, 2008a).

Anchor-based methods use some type of reference standard to determine the smallest important difference in a measurement instrument (Maughan & Lewis, 2010). Reference standards include global rating of improvement or predefined treatment goals (Ostelo, Deyo, Stratford, Waddell, Croft, Von Korff, Bouter, & de Vet, 2008a). An example of an anchor-based analysis is the Receiver Operating Characteristic (ROC) curve (Grotle et al., 2004).

### 3.3.1 Effect Size (ES) and Standardized Response Mean (SRM)

Effect size statistics measure the ability of a health measurement scale to detect a signal (change) among the noise (variability of the population) of a patient sample (Frost et al., 2008). It relates the magnitude of the change to the variability in the score (Cohen, 1977). The ES is calculated by dividing the mean change score of the patient population by the standard deviation of the baseline scores (Liang, 2000). Because ES is an estimation of a population parameter, it is not influenced by sample size (Norman & Streiner, 2008). If the mean change score is divided by the standard deviation of the change in scores between two assessment points, this is termed the standardized response mean (SRM) (Liang, 2000). Effect size calculations are most frequently used with a simple before and after single-group design (Stratford et al., 1996).

One disadvantage of the ES is that it assumes that all patients will change their scores in the same direction and imprecision is introduced if stable patient scores are included in the summary statistic (Liang, 2000). Also, a highly select group of patients who have almost the same baseline scores will inaccurately overestimate the ES (Liang, 2000).

The SRM is a variation on the ES. This statistic is defined as the ratio between the mean change in a single group to the standard deviation of the change scores (Norman & Streiner, 2008). As with ES, this standardized measure of change is used with a simple, single-group design (Stratford et al., 1996).

It is important to note that both ES and SRM evaluate the average change only. They assess the difference between the initial and the follow-up measurement and they cannot be used to assess a health status measures' capability to assess change to a varying degree as they assume that all patients have changed by the same amount (Stratford et al., 1996).

### 3.3.2 Paired T-Test

The t-test is a statistical test that is used to compare the mean change and is based on the ratio of the measured change to the standard error of the change (Norman & Streiner, 2008). This test is closely related to the SRM as it measures the difference between initial and follow-up scores. The t-test is calculated by multiplying the SRM by the square root of the number of subjects taking part in the study (Stratford et al., 1996). As with the ES and SRM statistics, the paired t-test is limited in that it can only evaluate change and not the extent to which a measure is capable of detecting different amounts of change. It is only used with single-group design studies (Stratford et al., 1996). Another limitation of the paired t-test is that it is influenced by sample size as the number of subjects is considered in the denominator of the equation (Norman & Streiner, 2008).

### 3.3.3 Receiver Operating Characteristic (ROC) Curve

ROC curve analysis is used when a reference standard is available to create two subgroups that change by different amounts (Norman & Streiner, 2008). This analysis can be used with any of the three types of multiple group designs (Stratford et al., 1996). An ROC curve is created by plotting sensitivity on the y-axis against 1-specificity on the x-axis (Stratford et al., 1996). Sensitivity is defined as the number of cases correctly identified by the outcome measure as experiencing an important change divided by all patients who truly underwent an important change (Stratford et al., 1996). Specificity, however, refers to the number of patients who were accurately identified by the outcome measure as not undergoing an important change divided by all patients who did not experience change (Stratford et al., 1996). The area under the curve (AUC) depicts the probability of accurately determining a patient who has undergone an important change in status (Stratford et al., 1996). An AUC of 0.50 indicates that a measure does no better than chance at distinguishing between the two groups of patients who changed by different amounts. The greater the area under the ROC curve, the greater the measure's ability to identify important change.

### 3.3.4 Analysis of Variance (ANOVA)

Analysis of variance is used when comparing two or more groups of patients (Stratford et al., 1996) and is appropriate for use with any of the multiple-group designs. An ANOVA calculation is performed by creating a sum of squares representing the differences between individual group means and another sum of squares representing the variation present between the groups (Norman & Streiner, 2008). The degrees of freedom, defined as the number of unique pieces of information, are necessary as each sum of squares (both between and within) is then divided by its degree of freedom, called the mean square. Finally, the mean square between is divided by the mean square within to create the F-ratio (Norman & Streiner, 2008). The F-ratio is then evaluated using a standardized table to determine if the results are statistically significant. The F-ratio, however, only determines if a difference amongst the means exists. It cannot determine where that difference is occurring and further comparisons of the means must then be performed to locate the statistically significant change (Domholdt, 2000).

#### *3.3.4.1 Norman's $S_{repeat}$*

Norman's $S_{repeat}$ is derived from a repeated-measures ANOVA. This coefficient represents the variance of the *group x time* interaction divided by the *group x time* interaction plus error variances (Stratford et al., 1996). This type of statistical analysis is most often used with all types of the multiple-group designs.

#### *3.3.4.2 Norman's $S_{ancova}$*

Norman's $S_{ancova}$ is derived from an analysis of covariance. This coefficient represents the group variance divided by the group plus error variance (Stratford et al., 1996). The dependent variable is the follow-up score and the covariate is the initial score. This type of analysis is most appropriate for the third type of multiple-group design in which a reference standard is used to assess changes in subjects’ status at the second follow-up assessment.

### 3.3.5 Correlation Coefficient

A correlation coefficient is used when it is expected that many patients will truly change by different amounts. Change scores for the outcome measure of interest are correlated with change scores from the reference standard. The greater the correlation, the more adept a measure is at detecting change (Stratford et al., 1996). Spearman’s rho correlations are used when both variables are ranked (Domholdt, 2000), which is appropriate for head-to-head-comparison studies using a global change index. Correlation coefficients assume that there is a linear relationship between the two variables and that each variable has enough variability to demonstrate a relationship (Domholdt, 2000).

## 3.4 Chapter Summary

This chapter provided a brief review of the various study designs and analytic techniques that are commonly used in outcome measure research to determine sensitivity to change or responsiveness. The advantages and limitations of each study design have been illuminated, highlighting the importance of choosing the appropriate design for a specific study. The common analytic techniques were briefly described and the conditions for their proper use were identified. It is important to consider both the study design type and the analysis technique to allow for optimal execution of an outcome measure investigation.

# Chapter 4: Ethical Considerations in Clinical Research

## 4.1 Introduction to Chapter

This thesis does not require ethics approval as no active participant recruitment and involvement was performed. The purpose of this chapter is to review the important and compulsory steps of gaining ethics approval that are necessary for conducting a head-to-head comparison study.

## 4.2 Research Ethics Board

Research ethics boards are designed to ensure that the principles of ethics, laid out by the Nuremberg Code and Canadian Law, are followed diligently by the investigators (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). Prior to the conduct of a study, any research protocol involving human subjects must be reviewed and approved by a research ethics board (REB) (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). This must occur before any participants are approached to join the study.

The main function of the REB is to consider the risk of harm versus the potential benefits of the research and how they would affect each individual participant (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). Ideally, the foreseeable harms should not outweigh the potential benefits and subjects should not be exposed to unnecessary risk of harm. It is the responsibility of the research team to ensure that the benefits associated with the research study are highlighted as much as possible while the risks are minimized. Throughout the duration of the research, continuing ethics review is essential, the frequency and depth of these reviews being proportional to the amount of risk to which each participant is exposed (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011).

If a research protocol is deemed to involve only minimal risk to the participants, the ethics review process involves much less scrutiny than studies in which risk to subjects is more than minimal.  Overall, the REBs are in place to protect human participants and the REBs have the authority to approve, disapprove, modify or terminate any proposed or ongoing research (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011).

Prior to initiating a head-to-head comparison study with human subjects comparing the Roland Morris Questionnaire and the Oswestry Disability Index, REB approval would need to be granted.  Any suggested modifications to the initial research protocol would need to be assessed and changed, while ongoing ethics approval throughout the research period would be required.

## 4.3 Informed Consent

Foremost to any research involving human subjects is the requirement of voluntary and informed consent from each participant.  Researchers must provide prospective subjects with adequate information about the nature of the research being conducted and the associated risks and benefits that may be encountered throughout the study (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011).  The process of informed consent is ongoing and begins at the initial meeting with each subject and continues for the length of the research.  Participants must be provided with adequate time to deliberate their potential involvement and subjects must also be made aware that they can withdraw from the study at any time without penalty. Consent must be voluntary and not be given under coercive or manipulative terms (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011).

In circumstances where a participant is unable to provide informed consent, such as with children or with people who are not deemed to be legally competent, informed consent must be obtained and maintained throughout the study timeframe from the authorized representative of the incompetent individuals (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). As the ability to be competent can change throughout the course of a study, such as with patients who are initially critically ill and unable to provide their own informed consent, researchers must demonstrate how they will continually assess this change in status and receive continued informed consent from the appropriate individual. Patients who are deemed legally incompetent can agree to participate in a research study but this may not be sufficient to allow entry into the study (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). However, if a person who is legally incompetent refuses to participate in a study this is sufficient to preclude them from the research.

Initial and continued voluntary informed consent is necessary from each participant in a head-to-head comparison study. It would be essential to inform each subject about the goals of the research, the nature of the involvement and the expected risks and benefits that may be inherent to participating. Individuals would be informed that should they agree to participate they can withdraw from the study at any time without penalty. The informed consent obtained would need to be received without coercion.

## 4.4 Privacy and Confidentiality

Privacy and confidentiality are recognized as fundamental human rights in the Canadian Charter of Rights and Freedoms and maintaining the privacy and confidentiality of both the personal and the research information of each participant is crucial throughout a study (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). Personal information is defined as "information relating to a reasonably identifiable person who has a reasonable expectation of privacy, including information about personal characteristics such as culture, age, religion and social status, as well as their life experience and educational, medical or employment histories" (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011)(p.3.2). Although no single item can be relied upon to identify a participant with absolute certainty, care must be taken to ensure that whichever piece(s) of information are used to identify a subject, that individual privacy is protected. Researchers are required to prevent any unauthorized releases of identifiable personal information.

It is the responsibility of the research team to maintain both the confidentiality and anonymity of each participant throughout the study period as well as after the completion of the study. Each participant determines when, how and to what extent any personal information is shared with others and this is often discussed as part of the informed consent process. Researchers are required to prevent any unauthorized releases of identifiable personal information. Safeguards such as using code numbers for each subject and for their personal information and using locked rooms or cabinets to store information are all means of protecting patient anonymity and confidentiality (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011).

Research Ethics Board approval is necessary for any research involving the use of questionnaires to collect patient information. Any threats to privacy, confidentiality and anonymity must be minimized and if the information attained will be used for secondary research, this must be addressed with each participant and consent obtained (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011).

Any patient information collected in a head-to-head comparison study would need to be protected through number coding and storing the information in a secured location. If the information gained from the research, both demographic and experimental, may be used for a secondary purpose, this must be explained to each subject and their consent obtained to allow for the subsequent sharing of the material.

## 4.5 Conflict of Interest

In research, the interests of the participants should take priority over the interests of others. However, a conflict of interest arises whenever the obligations of an investigator to research subjects are compromised or threatened by that individual's personal interests in accommodating financial arrangements, personal relationships, the institution's goals, job security or personal goals (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). Conflicts of interest create inequity and unfairness in the research process and can result in improper enrollment of study participants, reduced informed consent, bias in the results and possible reluctance in disseminating the research findings, all of which can occur in a head-to-head comparison study.

All researchers are required to disclose any actual, potential or perceived conflicts of interest to the REB, who will then analyze the potential implications of the stated conflicts of interest (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). Actions may include having the researchers disclose their conflicts of interest to each study participant or the researcher may be required to abandon one or all of the conflicts. Along with researchers, members of the REB and the institution through which the study is being run may also have conflicts of interest. Members of the REB who admit to possible conflicts of interest must withdraw from the board or an assessment will be undertaken to determine the risk of bias should that member remain a part of the REB (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011). The institution itself must provide resources to allow for the independent functioning of the REB which should enable the REB to act autonomously from the institution.

The inclusion of all persons in research is a fundamental ethical consideration. Historically, women, minority groups, older adults and individuals who are unable to give informed consent have often been excluded from research. It is essential that all individuals be allowed to participate in research so that a fair distribution of research benefits is permitted to all groups. Ethical research protocols now involve “balancing the need to protect vulnerable persons with the need to ensure that these groups have access to participation in potentially beneficial research” (*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS)* 2011)(p.5.2).

The recruitment process for a head-to-head comparison study must allow for the inclusion of all persons, including women, minority groups and those who cannot give informed consent.  Every individual who meets the inclusion criteria for the study has the right to be a part of research if they so choose.

## 4.6 Chapter Summary

Although this thesis did not require ethics consent, a brief review of the ethical considerations that are integral to the conduction of patient-related research was provided. Informed consent, privacy and confidentiality, conflicts of interest and the role of the Research Ethics Board are necessary and important aspects of ethics that must be addressed when conducting research to ensure fair, unbiased and inclusive results.

# Chapter 5: Methods

## 5.1 Introduction to Chapter

The following chapter details the methodological process for searching, retrieving and analyzing head-to-head comparison studies of the Roland Morris Questionnaire (RMQ) and the Oswestry Disability Index (ODI). The research question and purpose of the systematic review as well as the specific search terms and databases used are provided.

## 5.2 Purpose of Thesis

The primary purpose of this systematic review was to answer the following research question: Is there a difference in the sensitivity to change between the Roland Morris Questionnaire and the Oswestry Disability Index in their ability to measure pain-related functional status in persons with low back pain? A secondary purpose of this thesis was to create a quality criteria form that could be used to critique head-to-head comparison studies. This form was then used to assess articles comparing both the RMQ and the ODI. Finally, a head-to-head comparison of available RMQ and ODI data was performed to determine if there is a difference in the measures' sensitivity to change.

## 5.3 Search Strategy

A systematic review of the literature was initially performed on March 21, 2011 to locate articles comparing the psychometric properties of the RMQ and the ODI in patients with low back pain. An updated search was performed on July 17, 2011 to ensure no articles were missed in the initial search. Five electronic databases were selected and searched, including

Pubmed, EMBASE, OVID (Medline), CINAHL and AMED. Pubmed and EMBASE were searched for the period of 1980 to present, MEDLINE titles were reviewed from 1948 to July 2011 while AMED was searched from 1985 to present. Both the author and supervisor (A.N. and P.S.) performed these searches to reduce the likelihood of any key titles being missed in each database. Using the Boolean operators "AND" and "OR", the following search terms were used in various combinations to help identify relevant articles: "roland", "oswestry", "respons*", "ROC", "reliab*", "sensitivity", "SRM", "ES", "longitudinal" or "correlation". Searches were limited to articles published in English.

### 5.3.1 Inclusion and Exclusion Criteria

Inclusion criteria for this review were: articles published in English between the years of 1980 to the present, patient populations of adults aged 18 years or older with either acute or chronic low back pain with or without surgical intervention, studies using the 24-item RMQ and versions 1.0 or 2.0 of the ODI, cross-cultural adaptations of each outcome measure, studies that included patients who provided informed consent, prospective study designs with dependent populations (same patients taking each outcome measure at more than one assessment time). Exclusion criteria included: studies that used the 23- or 18-item RMQ, articles not published in English and patients who had low back pain as a result of malignancy, spinal fracture, infection, inflammatory disease or unstable neurological conditions.

Abstracts from all papers retrieved from the searches were read and examined for inclusion into the study. A study was included if it compared the RMQ and the ODI in terms of their abilities to measure patient change (i.e., responsiveness, sensitivity to change or longitudinal validity). The articles identified in the search were comprised of differing study designs, including prospective cohort studies, systematic and non-systematic reviews and cross-sectional designs. In this study, I was interested only in articles where the same patient sample was assessed with both the RMQ and the ODI at more than one time point. Apparent relevant abstracts in non-English journals were considered only if an English abstract was provided. We did not, however, encounter any articles published in another language but which had an English abstract.

Articles were eliminated from consideration in this study for a variety of reasons, including: cross-sectional design, studies published in a language other than English, articles that were establishing cross-cultural validity of the RMQ and ODI and articles that provided a review of various outcome measures without direct comparison between the RMQ and ODI. Both the author and the supervisor independently examined potential articles to ensure their relevance to the study. Once an article was located, its list of references was scanned for any studies that may have been missed through the initial search. Authors were then contacted via email, first by the thesis author and then two to three weeks later by the supervisor if no response had been received. As one of the articles was written by this author's supervisor, no email contact was necessary. Each author was asked in the correspondance to review an appended list of relevant titles and asked to include any other relevant articles that may have been inadvertently overlooked by the both the electronic and reference list searches. The contacted authors were asked to provide several items of information from their research, including their baseline and change scores for both the RMQ and the ODI and their reference standard scores. Patient demographic information to allow for a comparison between participants in different studies was also requested.

## 5.4 Quality Criteria

Along with the systematic review, the second intent of this thesis entailed the creation of a quality criteria form to analyze head-to-head comparison studies. This form, created jointly by this author and her supervisor, is provided in Appendix 3. This form was developed to assess key aspects of head-to-head comparison studies, such as patient demographics, study design, outcome measure details, analysis techniques and results. Each article was independently analyzed using this quality criteria form by two raters: the author and the supervisor. The results of each assessment were discussed by the two raters and agreement on the scoring was obtained. A third party rater was available if any disagreement in scoring between the two primary raters was not ratified. No disagreements occurred that required the use of the third independent rater.

## 5.5 Analysis

Because investigators have often reported multiple change coefficients that at times are based on conflicting assumptions concerning the change characteristics of the sample, I was guided by the work of Stratford and Riddle (2005) in determining the most appropriate coefficients for a given study. If the investigators did not declare the expected change characteristic of the sample and the reference standard had three or more response options of hierarchical structure (e.g., global rating of change [GRC]), I considered the most appropriate analysis to be a correlation between the GRC and the RMQ and ODI's change scores (Stratford & Riddle, 2005).

When investigators did not perform a formal head-to-head analysis of the difference in change coefficients for the RMQ and ODI, I attempted to conduct a comparison based on information published in the manuscript or additional information requested from the author. I calculated Spearman's rank order correlation coefficient between the GRC and RMQ and ODI change scores, and applied Meng's test (Meng, Rosenthal, & Rubin, 1992) for correlated data to evaluate the difference in coefficients. This test requires knowledge of the correlation between the RMQ and ODI's change scores. When this information was not available I estimated the correlation by pooling all available data provided by investigators responding to my request for raw data.

I also considered Receiver Operating Characteristic (ROC) curve analysis reported by investigators. For this analysis the area under the curve (AUC) represented the change coefficient. When the investigators did not formally compare the AUC between measures, and I was able to obtain additional information from the authors, I applied Delong's test (Delong, Delong, & Clarke-Pearson, 1988) for correlated data.

For all formal change coefficient comparisons my null hypothesis was there would be no difference in the RMQ and ODI change coefficients. All tests were 2-tailed and a difference was considered statistically significant if $p < 0.05$. Data were analyzed using SPSS V19 and STATA V10.1.

## 5.6 Chapter Summary

A systematic review of the literature was performed using five electronic databases and through reviewing the reference lists of the articles retrieved. These articles compared the RMQ and the ODI and assessed their sensitivity to change in adult patients with low back pain. A quality criteria form was created by this author and her supervisor to be used to critique head-to-head comparison studies of the RMQ and the ODI.

# Chapter 6: Results

## 6.1 Introduction to Chapter

This chapter outlines the results of the literature search and reviews the articles that qualified for the systematic review and quality assessment. A brief overview of the qualifying articles for this thesis, the common themes that emerged from the quality criteria appraisals and the results of the statistical analyses are presented.

## 6.2 Search Results

The initial search using the terms “roland” and “oswestry” returned 107 results in Pubmed and 128 in EMBASE, the most of any combination of the terms in the five databases. OVID (Medline), AMED and CINAHL returned 98, 38 and 55 results respectively using both “roland” and “oswestry” as the search terminology. Combining the terms “roland” and “oswestry” with “respons*” and/or “sensitivity” helped to reduce the number of results returned in each database. Of the numerous results that were returned through the initial searches, 20 articles met the initial criteria to be considered for further review and possible inclusion into the systematic review by both the author and the supervisor. These same 20 articles were located through both the Pubmed and EMBASE databases. Nineteen of the 20 were found through OVID (Medline) while AMED and CINAHL each returned only eight of the twenty identified papers. Articles were eliminated from further evaluation for a variety of reasons, including a lack of direct comparison of the RMQ and the ODI, studies with a cross-sectional design, clinical trials assessing the effectiveness of treatments without a direct comparison between the two measures, cross-cultural adaptations of the RMQ or the ODI or articles that were not published in English. Through the systematic searching of the five databases, the same nine articles were located, though not every database listed all of the nine. Appendix 4 demonstrates the total number of articles located in each database and the subsequent relevant titles from those sources.

Based on the stated eligibility criteria, nine articles were deemed to be appropriate for inclusion in the systematic review. No additional relevant studies were identified in the reference lists of these articles. Eight of the nine authors of the eligible studies were then contacted via email, first by the author and then two to three weeks later by the supervisor if no response had been given. As one of the articles was written by this author's supervisor, no email contact was necessary. Of the eight authors that were emailed, three responded to requests by this author (M. Grotle, M. Davidson and A. Mannion) and one responded to the supervisor's request (A. Beurskens). The data provided by these four authors was compiled with the data from Stratford et al (1994).

### 6.2.1 Article Summaries

Beurskens and colleagues conducted a prospective study to analyze the responsiveness of three outcome measures for low back pain, two of which were the Dutch versions of the RMQ and the ODI (Beurskens et al., 1996). A seven-point global perceived effect scale was used to estimate clinically important change. The patient sample was drawn from subjects who were previously enrolled in a randomized control trial on the efficacy of continuous motorized traction for low back pain. These patients had to have symptoms of low back pain for a minimum of six weeks. Eighty-one of the original 151 subjects were asked to complete the ODI, the RMQ and the global perceived effect scale at baseline and five weeks later. Responsiveness was assessed statistically by both effect size and by ROC curve analysis. Of the eighty-one patients, five were found to have deteriorated over the five week period and the authors chose to exclude them from the final analysis. My analysis of Beurskens et al's data revealed a statistically significant difference in favour of the RMQ for both the correlational analysis ($p < 0.001$) and Receiver Operating Characteristic (ROC) curve analysis ($p < 0.003$).

The purpose of Coelho et al's study was to assess the responsiveness of the Brazilian-Portuguese version of the ODI (Coelho, Siqueira, Ferreira, & Ferreira, 2008). Thirty subjects with at least a three month history of non-specific low back pain were selected from the physical therapy departments at two University clinics. The study sample completed the Brazilian-Portuguese versions of both the ODI and the RMQ at baseline and six weeks post physiotherapy treatment. Participants also completed a seven-level global perception of change Likert scale at the six week follow-up, which served as the reference standard of change. In this study, responsiveness was evaluated using both effect size and ROC curves. For the ROC evaluations, the area under the curve (AUC) was found to be 0.82 for the RMQ and 0.73 for the ODI. There was insufficient information to determine if there was a statistically significant difference between these two measures.

The aim of Davidson and Keating's study was to compare the reliability and responsiveness of the five most commonly used outcome measures for low back pain, which included the RMQ and the ODI (Davidson & Keating, 2002). In this prospective study, 207 patients were recruited from multiple sites, including the outpatient physiotherapy departments at three hospitals, three community health services and four private physiotherapy clinics. One hundred and one patients completed the English versions of the RMQ and the ODI at both time periods (baseline to the six week follow-up) as well as a seven-point global change scale that was included as the comparison standard at the six-week follow-up. Responsiveness was quantified in three ways, including standardized response means, ROC curves and by considering the proportion of patients who changed by at least as much as the MDC. My reanalysis of the results noted no significant difference between the correlation coefficients of the RMQ and the ODI with the global change index (p = 0.78). Similarly, the difference in AUCs for both outcome measures was not statistically significant for either measure (p = 0.85).

The purpose of Frost et al's prospective cohort study was to compare the responsiveness of the RMQ and the ODI v. 2.1 (Frost et al., 2008). The English versions of both questionnaires were used. The study sample included 286 subjects with a minimum of six weeks history of low back pain. Two-hundred and one participants completed the study and provided data at both baseline and at the 12 month follow-up. A global transitions rating scale was used as the reference standard of change. Frost et al. did not provide a formal comparison of the difference in correlation coefficients between the measures' change scores and the global transition scale's results. However, using an estimated correlation of 0.70 between the RMQ and ODI, this author found that the difference in correlation coefficients was not statistically significant (p = 0.07). For the ROC curve analysis, the point estimate of the AUC favoured the ODI, however insufficient information was available to determine whether this difference was statistically significant.

Grotle and colleagues' 2004 article compared the responsiveness of four functional outcome measures, including the Norwegian versions of the RMQ and the ODI (Grotle et al., 2004). The study population consisted of two samples of patients drawn from a specialist back clinic: an acute sample with a history of low back pain for less than three weeks and a chronic sample with a history of low back pain for greater than three months. Fifty-four patients with acute low back pain completed the RMQ and the ODI at baseline and at four weeks while 50 patients with chronic low back pain completed the outcome measures at baseline and three months later. A six-point global index of change was used to assess the subjects' overall change in their back pain since baseline. Grotle et al also utilized the expected clinical course of the low back pain as a means of evaluating responsiveness, with the assumption that the acute group would show greater improvement than the chronic sample. Complete data was available for 51 patients in the acute sample and 48 patients in the chronic sample. Spearman correlations coefficients were used to determine the correlations between the outcome measures and the reference standard and ROC curves were used to determine responsiveness. Based on a review of the data, no significant correlation was found between either the RMQ or the ODI in regards to the six-point global index of change (acute: p = 0.30; chronic: p = 0.12). There were also no statistically significant findings in the ROC curves for either scale in both groups (acute: p= 0.42; chronic: p = 0.23).

The primary goal of Kopec et al's study was to evaluate the measurement properties of the Quebec Back Pain Disability Scale (Kopec et al., 1995). However, this study also allowed the comparison of the ODI and RMQ's abilities to detect change by comparing their change scores to a retrospective 15-point global rating of change (i.e., the reference standard). The study sample was composed of patients from many sites including physiotherapy, physiatry, general practice, orthopaedic, pain, and rheumatology clinics. Change data were obtained from 178 patients and the interval between assessments was 6-months. Some patients completed French versions of the measures and other patients completed English versions of the measures, the exact numbers for the change comparison are not clear. The investigators anticipated that many patients would change by different amounts. Accordingly, the most appropriate change coefficient is the correlation between the measures' change scores and the reference standard's rating of change. The correlations with the reference standard were RMQ 0.47, ODI 0.35. Although the investigators did not formally compare this difference, I estimated it to be statistically significant (p = 0.018).

Mannion et al's 2006 study was a continuation of their 2005 article that was performed to develop a German version of the ODI (Mannion, Junge, Grob, Dvorak, & Fairbank, 2006). This study was conducted to determine the sensitivity to change of this newly validated German ODI. Both the German ODI and RMQ were administered to 63 patients undergoing planned spinal surgery, of which 57 completed the second follow-up questionnaires at six months. All subjects had been referred to the hospital's spine unit for a variety of spinal conditions, including spinal surgery, herniated disc, spondylolisthesis or degenerative disc disease with chronic low back pain. At the second follow-up, each subject also completed a six category Likert scale to determine how the subject's back/leg pain and disability had changed since the surgery. The correlation of the outcome measure scores with the global outcome scale and the ROC curve analysis were used to evaluate responsiveness. My review of the results found no significant difference between the correlation coefficients for the RMQ and the ODI (p = 0.62). There was no statistically significant difference between the AUC for the RMQ and the ODI (p = 0.75) as the point estimates for the ROC curves were very similar.

Maughan and Lewis (2010) conducted a prospective, single site study to determine the responsiveness of five outcome measures in patients with low back pain (Maughan & Lewis, 2010). The five questionnaires, including the RMQ and the ODI, were administered at baseline and after a five week back class intervention. Consecutive patients with low back pain symptoms of at least three months were referred to the study by their physiotherapist from the

Pulross clinic in Brixton. 63 subjects initially completed the English versions of the RMQ and the ODI at both baseline and follow-up, including a seven-point global impression of change scale at the five week assessment. Forty eight patients provided complete data at both time points. Responsiveness was measured by the standard error of the measurement (SEM) and by ROC analysis. My review of the results noted an AUC of 0.64 for the RMQ and 0.67 for the ODI. Unfortunately, due to a lack of statistical information, we were unable to determine if this difference was statistically significant.

Stratford et al performed a prospective cohort study to compare the ability of the RMQ, ODI and the Jan van Breeman Institute (JVB) pain and functional capacity questionnaire to detect change over time (Stratford et al., 1994). The original English versions of both the RMQ and the ODI were used with their scores being compared to two 15-point global rating scales to assess change. One scale assessed amount of change and the other considered the importance of change. Eighty-eight consecutive patients with a history of mechanical low back pain with a mean symptom duration of 48 days were recruited from the outpatient physiotherapy department at a large teaching hospital. Subjects were randomly assigned in a balanced block of three to complete the three questionnaires prior to beginning physiotherapy and then four to six weeks later. Of the 88 participants 74 had complete data on the RMQ, ODI, and global ratings of change. My reanalysis of the 74 participants with complete data demonstrated no significant difference between the RMQ and ODI for the correlational (p = 0.63) and ROC curve areas (p = 0.45).

## 6.3 Quality Criteria Results

The quality criteria form was created as a means of analyzing and appraising head-to-head comparison studies of competing outcome measures that assess the same attributes or condition.  Patient demographics, symptomology, study design, outcome measure description, analysis and results are all aspects of a study that are evaluated using this form.  It is the purpose of this form to provide a reader with the criteria necessary to determine the quality of a head-to-head comparison study in a variety of methodological and analytical areas.  This quality criteria form was used to assess the nine articles found that compare the RMQ and the ODI and to highlight the key aspects of each study.  My evaluation of each article is provided in Appendix 5.

### 6.3.1 Study Purpose and Subject Demographics

All nine articles clearly stated their research question which helps to direct the reader towards the content and purpose of the studies.  The majority of the articles (seven out of nine) also had well defined eligibility criteria, though only four of the articles commented on the location of their subjects' LBP.  Both Beurskens et al and Mannion et al did not state the inclusion or exclusion criteria explicitly.

### 6.3.2 Study Design

In terms of study design, most articles stated the type of study, such as a prospective cohort design, and the settings of the studies (such as hospital outpatient department or clinic).  Stratford et al were the only authors to describe the measurement conditions applied during their study which involved a block randomization pattern of the outcome measures.  I was unable to determine based on the limited information in the other eight articles if the RMQ and ODI were completed in a specified order by each subject.

Each of the nine articles stated the length of time between the baseline and follow-up assessments though only Davidson et al provided justification for the time between measurements, explaining that the six week follow-up was a commonly used timeline for reassessment of symptoms in a clinical setting (Davidson & Keating, 2002).

All nine articles identified a reference standard to use as an external evaluation of patient change over the course of the studies. These reference standards of change ranged from six to 15 point scales and were provided to the subjects at the follow-up assessments. Unfortunately, based on the self-report nature of these reference standards, they were not independent of the outcome measure. Only Grotle et al used both the expected clinical course as well as a global change index to assess each subject's change from baseline to follow-up.

### 6.3.3 Measure Description

Overall, six different translated versions of the RMQ and the ODI were used in the nine articles, with English being the most frequently used in five of the nine studies (Stratford et al, Kopec et al, Davidson et al, Frost et al and Maughan et al). French (Kopec et al,), Dutch (Beurskens et al), Norwegian (Grotle et al), German (Mannion et al) and Brazilian-Portuguese (Coelho et al) were the other five adaptations used. Each translation, as well as the original English versions, has been previously validated in patients with low back pain. Between the nine studies, two versions of both the RMQ and the ODI were used, including the original and the modified versions of the RMQ and versions 1.0 and 2.0 of the ODI. No noted modifications were made by any of the authors in their respective studies to either of the outcome measures.

### 6.3.4 Sample Size

Sample sizes ranged from 30 to 201 participants, although these figures represent the final sample size totals after patient drop-outs were calculated.  None of the authors performed specific sample size calculations, obtaining subjects from previous research studies or by convenience through clinics or outpatient departments.

### 6.3.5 Analyses

Eight of the nine authors utilized the AUC from the ROC characteristic analysis as one method to evaluate sensitivity to change of the ODI and the RMQ.  Spearman's rank correlation coefficients were used to estimate the correlations between the reference standard and the two instruments in four of the nine articles.  Where possible, I performed these comparison analyses if the data were available in the articles or through our contact with the authors.  Only two studies performed a formal comparison of the two measures (Stratford et al and Davidson et al). Grotle et al and Davidson et al were the only two authors who accounted for the dependent nature of the data in their formal analysis, with Grotle et al using the expected clinical course of low back pain as a second reference standard and Davidson et al using the natural clinical course as a construct for change.

### 6.3.6 Quality Results Summary

Descriptive statistics of age, gender and the duration of back pain were consistently provided by the authors; however none of the authors stated the subjects' co-morbidities and just over half of the nine studies listed the subjects' work and pain pattern distributions.  The proportion of unanswered or duplicated responses was stated by four of the nine articles while only Beurskens et al and Davidson et al commented on patient losses to follow-up.

Overall, the majority of the studies provided descriptive statistics for each measure in terms of subjects' pre-scores, post-scores and change scores.  Only Kopec et al did not provide this data for the RMQ and the ODI.  P-values or confidence intervals were provided for the individual outcome measures' change statistics by six of the nine authors while four of the nine articles presented the p-values or confidence intervals for the between measure comparisons.

## 6.4 Comparison of Results

Using the data provided in each of the articles or by the five authors who replied to our email requests, I was able to report the head-to-head comparison results of the RMQ and the ODI or perform the appropriate calculations to determine the correlation between these two outcome measures.  Table 6-1 presents the Spearman correlation coefficients between the reference standard (global rating of change) and the RMQ and the ODI change scores.  Both the Z-scores and the p-values are provided, with a p-value of less than 0.05 as the standard for statistical significance.

Based on the calculations provided by the authors or by my analyses of the data, only Beurskens et al and Kopec et al had a statistically significant difference between the ODI and the RMQ change scores and their correlation to the reference standard.  In Beurskens et al's study, a Z-score of 3.28 with a corresponding p-value of 0.001 was noted in favour of the RMQ.  Kopec et al's data showed a Z-score of 2.36 with an associated p-value of 0.018, also in favour of the RMQ.  All other calculations did not reach statistical significance, as seen in Table 6-1.  I was unable to calculate the correlation coefficients for Coelho et al and Maughan et al as there was insufficient information provided and we did not receive a response from our data request.

**Table 6-1: Correlation Coefficient Comparison**

| **Study** | **Roland-Morris** | **Oswestry** | **Difference (Z, p)** |
|---|---|---|---|
| **Beurskens (n = 76)** | 0.72 | 0.46 | 3.28, 0.001 |
| **Davidson (n = 101)** | 0.49 | 0.51 | 0.27, 0.78 |
| **Frost (n = 201)** | 0.38 | 0.47 | 1.90, 0.06* |
| **Grotle (all data, n = 99)** | 0.75 | 0.68 | 1.64, 0.10 |
| **Grotle (acute, n = 51)** | 0.74 | 0.67 | 1.04, 0.30 |
| **Grotle (chronic, n = 48)** | 0.61 | 0.49 | 1.56, 0.12 |
| **Kopec (n = 178)** | 0.47 | 0.35 | 2.36, 0.018* |
| **Mannion (n = 57)** | 0.67 | 0.69 | 0.49, 0.62 |
| **Stratford (n = 74)** | 0.56 | 0.53 | 0.48, 0.63 |

*Based on an estimated correlation of 0.72 between the Roland and Oswestry change scores

Table 6-2 presents the AUC results of the ROC analyses. I was able to obtain the ROC area comparisons for five of the nine studies. I had no available data from Kopec et al and I was unable to calculate the AUC for Coelho et al, Frost et al and Maughan et al with the data provided. Through the application of Delong's test, the AUC's of both the RMQ and the ODI were compared to each other. Of the eight articles for which I had data, only Beurskens et al were noted to have a statistically significant comparison, with a $X^2_1$ score of 8.58 and a p-value of 0.003, in favour of the RMQ.

**Table 6-2: Receiver Operating Curve Area Comparison**

| Study | Roland-Morris | Oswestry | Difference( $\chi_1^2$, p) |
|---|---|---|---|
| **Beurskens (n = 76)** | 0.93 | 0.76 | 8.58, 0.003 |
| **Coelho (n = 30)** | 0.82 | 0.73 | not available |
| **Davidson (n = 101)** | 0.76 | 0.77 | 0.04, 0.85 |
| **Frost (n = 201)** | 0.69 | 0.75 | not available |
| **Grotle (GRC all data, n = 99)** | 0.89 | 0.84 | 1.95, 0.16 |
| **Grotle (GRC acute, n = 51)** | 0.93 | 0.87 | 0.66, 0.42 |
| **Grotle (GRC chronic, n = 48)** | 0.83 | 0.75 | 0.23, 0.23 |
| **Grotle (acute chronic)** | 0.73 | 0.71 | 0.29, 0.59 |
| **Mannion (n = 57)** | 0.84 | 0.85 | 0.11, 0.75 |
| **Maughan  (n = 48)** | 0.64 | 0.67 | not available |
| **Stratford (n = 74)** | 0.85 | 0.82 | 0.58, 0.45 |

Because of dissimilarities in study design, patient characteristics and conditions of measurement, no attempt was made to perform a meta-analysis.

## 6.5 Chapter Summary

The systematic review of the literature revealed nine articles that met the inclusion and exclusion criteria for this thesis. The authors of these nine articles were contacted and we received responses and data files from five of the nine authors. This chapter presents the results of the quality criteria evaluations and the correlational and ROC results, which demonstrated a small but significant difference between the RMQ and the ODI in favour of the RMQ.

# Chapter 7: Discussion

## 7.1 Chapter Introduction

Low back pain (LBP) is a frequent and persistent cause of disability and is estimated to affect 80 to 85 percent of individuals at some point in their lifetime (Hoy et al., 2010). Due to the high prevalence of LBP in society, there has been an oversaturation of outcome measures created to assess and interpret LPB impairments and activity limitations. Two of the most commonly used LBP outcome measures are the Oswestry Disability Index (ODI) and the Roland Morris Questionnaire (RMQ) (Roland and Fairbank 2000). Due to this abundance of competing outcome measures, the primary intent of this thesis was to perform a systematic review of the literature to identify articles that compared the ODI and the RMQ and to determine if there was a difference in the sensitivity to change of these two outcome measures in their ability to quantify pain-related functional status in persons with LBP. A secondary intent of the thesis was to create a quality criteria form to be used to evaluate head-to-head comparison studies and then apply these criteria to the articles identified by the systematic review. This final chapter will discuss these results in more detail and provide recommendations regarding the methodological, analytical and design considerations concerning head-to-head comparison studies of competing outcome measures.

## 7.2 Literature Search

A comprehensive and systematic investigation of the literature was first performed in March 2011 to locate all articles that compared the RMQ and the ODI in terms of their responsiveness or sensitivity to change. To ensure that no new relevant articles were added to the databases and to confirm that no titles had been missed in the initial search, both this writer and her supervisor performed multiple systematic searches of the databases, with the most recent search being performed on July 17, 2011. In total, five databases were selected and searched, including Pubmed, EMBASE, AMED, OVID (Medline) and CINAHL. These specific databases were chosen based on their inclusion of premier allied health research (AMED and CINAHL), their comprehensive list of titles (Medline and Pubmed) and for their inclusion of both European and international research (EMBASE). By rigorously searching these five databases, the likelihood of key articles being missed was greatly reduced. As well, given that two separate individuals searched these databases multiple times beginning in March 2011, we have accounted for the possibility of newly published research that may have been missed in the first literature search. In addition, the chances of overlooking important studies is minimized by two separate individuals performing these searches independently of one another.

Of the five databases used to locate relevant titles, both Pubmed and EMBASE returned the most number of articles, 107 and 128 respectively, and resulted in the most relevant titles through the combination of “roland” and “oswestry” as the search terms. The search results of these two databases both contained the 20 articles that met our initial criteria to allow for further review. OVID (Medline), CINAHL and AMED located 98, 55 and 38 articles respectively using the same two terms, while including only 19, 8 and 8 of the original 20 studies. When combining “roland” and “oswestry” with “respons*” and/or “sensitivity” in Pubmed and EMBASE, the number of results dropped substantially. In all, Pubmed, EMBASE and OVID (Medline) resulted in locating all nine relevant articles that were eventually included in this systematic review. Although AMED and CINAHL also located a varying combination of the nine, neither of these two databases contained all of the included titles. Overall, using Pubmed, EMBASE and OVID (Medline) resulted in the highest number of returns as well as each containing the nine eligible studies and were the most productive databases of the five. Combining “roland” and “oswestry” using the Boolean operator “and” produced the greatest number of citations and facilitated the discovery of the nine relevant articles. Although helping to streamline the results of the literature search, using the terms “respons*” and “sensitivity” resulted in missing several of the nine key titles and therefore less helpful than simply using “roland” and “oswestry”.

I used the nine titles that were found through the literature search as potential sources of additional studies by reviewing each reference list that accompanied the articles. Both this writer and the supervisor reviewed each of the reference lists as a means of locating other relevant studies, although these perusals did not result in any other eligible head-to-head comparison studies of the RMQ and the ODI. Along with our request for both patient demographic and statistical data, we asked each of the eight authors to review our appended list of 20 articles and to suggest any other potentially relevant title. Of the five authors who responded with data, none suggested any further head-to-head comparison studies for me to review, which supports the comprehensive nature of my systematic review.

Due to a thorough literature search, which included using five comprehensive databases and two independent searchers, email contacts with the five responding authors and a review of the reference lists of the nine qualifying head-to-head comparison studies, it is unlikely that a comparison study of the RMQ and the ODI was overlooked for this thesis.

## 7.3 Quality Criteria Summary

There were several areas of deficiency that became apparent upon review of the nine articles using the quality criteria form and these deficiencies limit the interpretation of the studies.

All nine selected articles had a clearly defined purpose statement or research question, allowing a reader to discern the direction of the research and to determine if the study is applicable to their clinical interest.

Although the majority of the articles had clearly defined eligibility criteria, two authors, Beurskens et al and Mannion et al, did not state the inclusion or the exclusion criteria explicitly. This makes it difficult to determine the type of patient population involved in the research. Having an understanding of the population demographics allows for the greater generalizability of the evidence to a clinical population. It is recommended that all head-to-head comparison studies list their eligibility criteria to clarify the type of subjects included and excluded from the research.

Overall, the majority of authors stated the study design explicitly, although Kopec et al, Mannion et al, and Coelho et al failed to provide a clear statement of their study format. Head-to-head comparison studies require a prospective design with a minimum of two measurement points for each patient. Ambiguity in study design can compromise the interpretation of the results. Unfortunately, only Stratford et al provided information as to the measurement conditions of the RMQ and the ODI. A lack of standardization in the order of administration of the outcome measures could lead to bias in patient responses due to rumination by the patients (i.e., reflection on the questionnaire administered first may affect responses on the second questionnaire). Describing the specific order of measure administration would eliminate this bias from the research, although this may be difficult in studies that involve mailing the outcome measures to the subjects at the final follow-up as there is no means to ensure that each patient completes the forms in the same order.

Only Davidson et al justified why they chose the length of time between assessments, stating that their choice of a six week follow-up was similar to the length of time used in most clinical settings to reassess a patient’s low back pain symptoms. No other author validated their choice of follow-up timeline. I believe this explanation is important as clinicians may interpret the results of the head-to-head comparison differently if the reassessment timeframes are more in line with their clinical practice. Basing their follow-up on what they believe to be a commonly used clinical timeline improves the justification for clinicians to implement the results of a study.

Only Kopec et al and Grotle et al stated their expectation of their samples’ change characteristics. By acknowledging their beliefs prior to beginning the study there is less chance of interpretation bias of the results by the investigators since they have already stated the direction of their assumptions.

Only Grotle et al considered using an independent reference standard (expected clinical course) to assess baseline to follow-up change in their subjects. All other authors used a global change index (GCI) to measure change. Using the expected clinical course of low back pain as a reference standard reduces the bias inherent in the self-report GCI scales as it would be expected based on clinical presentation that patients with acute LBP would see greater improvement over the course of a study than those with chronic LBP. Grotle et al did state that they expected to see a larger improvement in scoring on the RMQ and the ODI in their subjects with acute LBP versus those who complained of chronic LBP. Using an independent reference standard helps to reduce both subject and evaluator bias and it is recommended that head-to-head comparison studies consider using this type of reference standard along with a GCI.

In terms of outcome measure descriptions provided by the authors, the majority of authors stated both the version and the language of the RMQ and the ODI used in their study. Since the various versions of the RMQ and the ODI have been validated individually and possibly with different populations, authors are encouraged to explicitly state the version of the outcome measure used so that the results of the head-to-head comparison study will clearly support the use of that specific version of the outcome measure in question.

None of the authors stated what they considered to be an important between measure difference in sensitivity to change, nor did they perform a sample size calculation prior to beginning their studies.  Accordingly, I cannot determine whether the observed differences in sensitivity to change coefficients represent meaningful and statistically significant change had the appropriate sample size been applied.  In addition, I cannot determine whether the studies were under-powered.

The appropriate analysis techniques, including ROC characteristics, were utilized in eight of the nine studies.  This type of statistical analysis is consistent with the expected change characteristics of the sample population.  Kopec et al did not apply an ROC analysis, but rather chose to report Norman's S sensitivity to change coefficient.  Only Stratford et al and Davidson et al attempted to compare the AUCs of the RMQ and the ODI to evaluate the differences in the abilities of the measures to analyze patient change.  By performing a formal comparison of the outcome measures the authors can draw more specific conclusions from their results as to which measure may be best able to assess a change in symptom status or clinical presentation. All future head-to-head comparison studies should attempt to compare the measures directly as this can promote a potential change in the favoured outcome measure used in a clinical or research setting as the measure which assesses patient change more accurately should be chosen more consistently than the comparison measure. Spearman correlations were performed by four of the authors (Frost et al, Grotle et al, Mannion et al and Stratford et al) to analyze the relationship between the global index of change and the change characteristics of the RMQ and the ODI. The more highly correlated the change scores are with the patients' perception of their change in status, the more useful the outcome measure will be at monitoring both improvements and worsening of symptoms.

Descriptive statistics of age, gender and the duration of symptoms prior to study entry were generally well done by all nine articles, although a list of co-morbidities was not provided by any of the authors. By providing extensive patient demographic information, the research can then be more appropriately applied to a similar clinical population. The more similar a study sample is to a clinical population, the more likely that the results will be useful in the clinical arena. Listing subject co-morbidities can also help to determine if the sample is comparable enough to the patients being seen by a treating therapist.

Four of the nine study authors (Beurskens et al, Kopec et al, Grotle et al and Stratford et al) commented on the number of incorrectly answered or missing responses from the outcome measures. Incorrectly scored forms could result in either an underestimation or an overestimation of patient symptom improvement which could bias the statistical results and the correlation between the external standard and the outcome measures. None of the authors performed an intention to treat analysis to account for any of the patient losses during their study timelines. It is possible that the results of our statistical re-analysis may be skewed by the lack of inclusion of all baseline patient data.

Descriptive statistics were presented in the majority of the nine articles, which provided an opportunity to see the score changes from baseline to follow-up. This allows for a visual representation of the changes made by a study sample in the units of the outcome measures. A summary of the baseline RMQ and ODI scores are provided in table 7-1. I was unable to determine the baseline scores in Kopec et al as there was insufficient information given in the article and we did not receive response to our email requests for data. It must be noted, however, that the purpose of Kopec's study was to investigate the Quebec Back Pain Disability Scale and not to perform a direct comparison between the RMQ and the ODI.

**Table 7-1: Baseline Roland Morris Questionnaire and Oswestry Disability Index Scores**

| Study | Roland-Morris | Oswestry |
|---|---|---|
| **Beurskens (non-improved: n = 38)** | 11.8 (5.1) | 29.1 (15.2) |
| **(improved: n = 38)** | 12.1 (4.7) | 26.2 (13.5) |
| **Coelho (all: n = 30)** | 11.1 (5.7) | 32.8 (18.9) |
| **Davidson (unchanged n = 47)** | 9.0 (5.2) | 35.0 (15.0) |
| **(improved: n = 52)** | 9.5 (5.9) | 35.0 (17.0) |
| **Frost (worse: n = 16)** | 7.3 (4.5) | 28.0 (12.5) |
| **(same: n = 76)** | 6.4 (4.3) | 22.6 (11.6) |
| **(better: n = 109)** | 4.9 (3.8) | 18.7 (9.4) |
| **Kopec** | Not available | |
| **Grotle (acute unchanged: n =11 )** | 7.6 (5.1) | 28.9 (18.0) |
| **(acute improved: n = 35 )** | 10.2 (5.0) | 29.6 (15.0) |
| **(chronic unchanged: n =20 )** | 9.2 (4.6) | 31.5 (12.2) |
| **(chronic improved: n =20 )** | 10.0 (4.0) | 30.0 (10.0) |
| **Mannion (all: n = 57)** | 15.0 (4.4) | 45.0 (15.6) |
| **Maughan (no change: n = 25)** | 14.0 (5.4) | 35.0 (20.2) |
| **(improved: n = 23)** | 9.0 (6.1) | 24.0 (18.2) |
| **Stratford (n = 74)** | 11.8 (6.2) | 40.5 (17.8) |

Six of the nine articles supplied p-values and confidence intervals for the individual measure change scores, which demonstrate whether or not a significant change in the score has occurred from baseline. Only four authors presented the between measure comparison statistics with p-values or confidence intervals. These values indicate the relationship between the RMQ and the ODI and whether or not there was a statistically significant difference between the two measures. Properly stated p-values and confidence intervals clearly inform a reader as to whether or not the evidence was found to be statistically significant and thus readers can then more readily decide whether or not to implement the research findings into their practice.

There were many deficiencies found through the evaluation of the nine head-to-head comparison studies with the quality criteria form created by this writer and her supervisor. The quality criteria form assists in highlighting the demographic, methodological and analytical considerations that should be reflected upon by every author who wishes to perform a head-to-head comparison of any outcome measures with a similar purpose. Ideally, all criteria would be explicitly stated for all head-to-head comparison studies. This is important as it allows other readers and future authors to have a clear understanding of the quality of the evidence available. Clinicians are then more likely to implement the evidence if the quality of the article and its results are clearly delineated. Researchers will also be more likely to advance the evidence, as opposed to duplicating it, if studies present all key methodological and statistical information. It is important to note, however, that some authors may be limited in the detail they can present depending on the publication restrictions of the academic journal to which they are submitting.

## 7.4 Head-to Head Comparison Results

My review and analyses of the available data from the nine eligible studies noted statistically significant differences in the Spearman correlation coefficient between the GRI and the RMQ and the ODI in both Beurskens et al ($Z = 3.28$, $p = 0.001$) and Kopec et al ($Z = 2.36$, $p = 0.018$), with the RMQ change scores attaining a strong correlation with the global change indexes. These results, however, provide only limited support for the RMQ as no other correlation coefficient attained statistical significance in the other articles and I was unable to calculate this statistic for Maughan et al and Coelho et al. These results must be interpreted with caution pending further methodologically sound inquiry.

Of the five ROC curve area comparisons for which I had data, only Beurskens et al's results had a statistically significant difference between the AUCs for the RMQ and the ODI, with a $X^2_1$ score of 8.58 and a p-value of 0.003. This indicates that the RMQ may be more accurate at predicting which subject's have had an important change in status versus the ODI; however it was the only study to demonstrate statistical significance between the AUC's for the RMQ and the ODI.

Only Grotle et al compared two groups of patients using an independent reference standard: one group with acute LBP and the other with chronic LBP. Grotle's results were not statistically significant. As this was the only group who considered this difference, I am unable to state which of the two measures is more appropriate to use with either clinical presentation.

In 2000, Roland and Fairbanks, in a narrative review, suggested that the RMQ may be better suited for patients with mild to moderate LBP disability while the ODI may be best utilized in situations where patients have persistent and severe LBP symptoms. They based these statements on the ability of the ODI to detect change even at high levels of disability (less ceiling effect) as opposed to the RMQ which is able to discriminate between patients even when ODI scores are at a minimum (less floor effect) (Roland and Fairbanks 2000).

Although the results are quite limited in this systematic review of head-to-head comparison studies, it is important to note that this is the first systematic review that has attempted to critically compare the sensitivity to change of both RMQ and the ODI. Other authors have conducted systematic reviews of the psychometric properties of the RMQ and the ODI, such as Davies and Nitz (2009) and Bombardier et al (2001). These authors however, did not formally compare the two measures' performance on the same patients and simply reviewed the various forms of reliability and validity of the two measures. Davies and Nitz only considered the English version of the outcome measures whereas this review included several translated and cross culturally adapted versions of the RMQ and the ODI (Davies & Nitz, 2009). Bombardier et al performed a systematic review to determine the minimum clinically important difference of the two measures and also did not formally compare the abilities of the RMQ and the ODI to measure change (C. Bombardier, Hayden, & Beaton, 2001). Neither author used a Spearman correlation or an ROC analysis to assess sensitivity to change.

Based on the available data, a slight advantage was noted in this head-to-head comparison study in terms of both the Spearman correlation coefficient and the ROC curve analysis in favour of the RMQ. However, all other statistical calculations did not achieve statistical significance and these insignificant calculations were far more numerous than the three that supported the RMQ over the ODI in terms of its association to the global change indexes and its AUCs. Further research is necessary before these results can be stated with absolute confidence.

Using the results of this systematic review to guide clinical practice, it would be recommended that the RMQ be used in favour of the ODI to assess the pain-related functional status of patients with low back pain, although this evidence is limited. Since no significant difference was noted between patients with acute or chronic LBP, I am unable to recommend which of the two outcome measures would be best utilized by these specific subgroups in a clinical setting. However, Roland and Fairbanks suggest that the RMQ be used in patients with lower levels of LBP-related disability while the ODI is recommended for use in patients with higher levels of disability (Roland & Fairbank, 2000). In terms of ease of use, the RMQ can be completed and scored in five minutes, which in a busy clinical practice may lend further advantage to the RMQ over the ODI.

## 7.5 Study Limitations

There are several limitations to both this systematic review and the newly developed head-to-head quality criteria form. First, it is possible that relevant titles could have been missed through our literature search. Although five databases were reviewed several times and by two independent searchers, there is the chance that other articles were overlooked. It is also possible that those four authors who did not respond to our email requests (Coelho et al, Frost et al, Kopec et al and Maughan et al) may have been aware of key titles that were missed and not included in this systematic review. It may also be possible that articles written in a different language may not have been included in the citation catalogue of the five databases, although this is unlikely considering our inclusion of EMBASE in our search, known for containing a large quantity of both European and international titles.

Since our quality criteria form was newly designed and created by this writer and her supervisor, it has yet to be utilized outside this thesis to assess other head-to-head outcome measure studies. Further use of this form with other comparable outcome measures is necessary to validate its ability to accurately assess a study's quality. As well, once this form is used on a more frequent basis, other researchers or clinicians may provide additional insights into other criteria that need to be included or removed from this form.

Another possible limitation is the varied study sample demographics between all nine articles. For example, some authors permitted subjects into their study with a history of spinal surgery while this was an exclusion criterion for other authors. This may have influenced our results since a lack of patient homogeneity can skew the data.

Because we could not directly ascertain the correlation between RMQ and ODI scores for Kopec's and Frost's data, we applied a pooled estimate of 0.72 based on data from the five investigators who provided raw data. We do not know the extent to which this estimate accurately reflects the actual correlation between these measures for Kopec's and Frost's samples. Also, it is possible that significant differences between the RMQ and the ODI could exist for other studies (e.g., Coelho and Frost) had sufficient information been available to conduct formal analyses.

It is also possible that the results of this study could be influenced by the pre-conceived ideas of this writer and her supervisor concerning the RMQ and the ODI based on our previous experiences with these two outcome measures. By basing our conclusions on the available statistical evidence, this should reduce the likelihood of bias or personal opinion influencing our results.

## 7.6 Chapter Conclusion

A systematic review of head-to-head comparison studies that compared the RMQ and the ODI was performed using five comprehensive databases. Nine articles met our inclusion criteria and these studies were evaluated using a newly-designed quality criteria form specific for head-to-head comparison studies. The form highlighted several common areas of inadequacy in terms of patient demographics, study methodology and data analyses.

Suggestions were made by this writer on the importance of improving these deficiencies. Based on the head-to-head comparison of the RMQ and the ODI performed for this thesis, the RMQ has shown a slight advantage over the ODI in its ability to detect changes in pain-related functional status in patients with LBP. Although these results are limited, it is the recommendation of this writer that based on the results of the head-to-head comparison, the RMQ be used to measure changes in patients with LBP. It is acknowledged that the current evidence is limited and future research is recommended.

# References

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*(1), 71-76.

Bombardier, C. (2000). Outcome assessments in the evaluation of treatment of spinal disorders: Summary and general recommendations. *Spine, 25*, 2003-2013.

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.

Cieza, A., Stucki, G., Weigl, M., Disler, P., Jackel, W., van der Linden, S., de Bie, R. (2004). ICF core sets for low back pain. *Journal of Rehabilitation Medicine : Official Journal of the UEMS European Board of Physical and Rehabilitation Medicine, (44 Suppl)*, 69-74.

Cleland, J., Gillani, R., Bienen, E. J., & Sadosky, A. (2011). Assessing dimensionality and responsiveness of outcomes measures for patients with low back pain. *Pain Practice : The Official Journal of World Institute of Pain, 11*(1), 57-69.

Coelho, R. A., Siqueira, F. B., Ferreira, P. H., & Ferreira, M. L. (2008). Responsiveness of the Brazilian-Portuguese version of the Oswestry Disability Index in subjects with low back pain. *European Spine Journal : Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 17*(8), 1101-1106.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd edition ed., pp. 443-507). Washington, DC: American Council on Education.

Davidson, M., & Keating, J. L. (2002). A comparison of five low back disability questionnaires: Reliability and responsiveness. *Physical Therapy, 82*(1), 8-24.

Davies, C. C., & Nitz, A. J. (2009). Psychometric properties of the Roland-Morris Disability Questionnaire compared to the Oswestry Disability Index: A systematic review. *Physical Therapy Reviews, 14*(6), 399-408.

Delong, E. R., Delong, D. M., & Clarke-Pearson, D. L. (1988). Comparing areas under two or more correlated receiver operating curves: A nonparametric approach. *Biometrics, 44*, 837-845.

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

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 Diseases, 39*(11), 897-906.

Domholdt, E. (2000). *Physical therapy research: Principles and applications* (2nd edition ed.). Philadelphia: Saunders.

Fairbank, J. C. T., Davies, J. B., Couper, J., & O'Brien, J. P. (1980). The Oswestry Low Back Pain Disability Questionnaire. *Physiotherapy, 66*(8), 271-273.

Fairbank, J. C., & Pynsent, P. B. (2000). The Oswestry Disability Index. *Spine, 25*(22), 2940-52; discussion 2952.

Finch, E., Brooks, D., Stratford, P. W., & Mayo, N. E. (2002). *Physical rehabilitation outcome measures: A guide to enhanced clinical decision making* (2nd edition ed.). Hamilton, ON: BC Decker Inc.

Frost, H., Lamb, S. E., & Stewart-Brown, S. (2008). Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index v2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. *Spine, 33*(22), 2450-7.

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*(21), E492-501.

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*(1), 130-140.

Guyatt, G. H., Deyo, R. A., Charlson, M., Levine, M. N., & Mitchell, A. (1989). Responsiveness and validity in health status measurement: A clarification. *Journal of Clinical Epidemiology, 42*(5), 403-408.

Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 1*(1), 73-75.

Hoy, D., Brooks, P., Blyth, F., & Buchbinder, R. (2010). The epidemiology of low back pain. *Best Practice & Research.Clinical Rheumatology, 24*(6), 769-781.

*Introductory tutorial for the tri-council policy statement: Ethical conduct for research involving humans (TCPS).* (2011). Retrieved April 3, 2011, from http://www.pre.ethics.gc.ca.libaccess.lib.mcmaster.ca/english/tutorial/

Jensen, M. P., Strom, S. E., Turner, J. A., & Romano, J. M. (1992). Validity of the Sickness Impact Profile Roland Scale as a measure of dysfunction in chronic pain patients. *Pain, 50*(2), 157-162.

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*, 139-143.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38*(1), 27-36.

Kopec, J. A., Esdaile, J. M., Abrahamowicz, M., Abenhaim, L., Wood-Dauphinee, S., Lamping, D. L., & Williams, J. I. (1995). The Quebec Back Pain Disability Scale: measurement properties. *Spine, 20*(3), 341-352.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*(11), 1183-1192.

Liang, M. H. (2000). Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. *Medical Care, 38*(9 Suppl), II84-90.

Mannion, A. F., Junge, A., Grob, D., Dvorak, J., & Fairbank, J. C. (2006). Development of a German version of the Oswestry Disability Index. part 2: Sensitivity to change after spinal surgery. *European Spine Journal : Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 15*(1), 66-73.

Maughan, E. F., & Lewis, J. S. (2010). Outcome measures in chronic low back pain. *European Spine Journal : Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 19*(9), 1484-1494.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlational coefficients. *Psychological Bulletin, 111*, 172-175.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 33-45) Psychology Press.

Norman, G. R., & Streiner, D. L. (2008). *Biostatistics: The bare essentials* (3rd edition ed.). Shelton, CT: People's Medical Publishing House.

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

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 I: Development of a reliable and sensitive measure of disability in low-back pain. *Spine, 8*(2), 141-144.

Shepard, L. A. (1993). Evaluating test validity. In *Review of research in education* (pp. 405-450) American Educational Research Association.

Shultz, K. S., Riggs, M. L., & Kottke, J. L. (1998). The need for an evolving concept of validity in industrial and personnel psychology: Psychometric, legal and emerging issues. *Current Psychology, 17*(4), 265-286.

Stratford, P. W., Binkley, F. M., & Riddle, D. L. (1996). Health status measures: Strategies and analytic methods for assessing change scores. *Physical Therapy, 76*(10), 1109-1123.

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., & Riddle, D. L. (2005). Assessing sensitivity to change: Choosing the appropriate change coefficient. *Health and Quality of Life Outcomes, 3*, 23.

World Health Organization. (2011). *International classification of functioning, disability and health (ICF).* Retrieved March 13, 2011, from http://www.who.int/classifications/icf/en

## Appendix 1: Oswestry Disability Index 2.0

Could you please complete this questionnaire. It is designed to give us information as to how your back (or leg) trouble has affected your ability to manage in everyday life.

Please answer *every section*. Mark *one box only* in each section that most closely describes you *today*.

**Section 1: Pain Intensity**
□ I have no pain at the moment.
□ The pain is very mild at the moment.
□ The pain is moderate at the moment.
□ The pain is fairly severe at the moment.
□ The pain is very severe at the moment.
□ The pain is the worst imaginable at the moment.

**Section 2: Personal care (washing, dressing, etc.)**
□ I can look after myself norm ally without causing extra pain.
□ I can look after myself normally but it is very painful.
□ It is painful to look after myself and I am slow and careful.
□ I need some help but manage most of my personal care.
□ I need help every day in most aspects of self care.
□ I do not get dressed, wash with difficulty, and stay in bed.

**Section 3: Lifting**
□ I can lift heavy weights without extra pain.
□ I can lift heavy weights but it gives extra pain.
□ Pain prevents me from lifting heavy weights off the floor but I can manage if they are conveniently positioned (e.g., on a table).
□ Pain prevents me from lifting heavy weights but I can manage light to medium weights if they are conveniently positioned.
□ I can lift only very light weights.
□ I cannot lift or carry anything at all.

**Section 4: Walking**
□ Pain does not prevent me walking any distance.
□ Pain prevents me walking more than 1 mile.
□ Pain prevents me walking more than a quarter of a mile.
□ Pain prevents me walking more than 100 yards.

□ I can only walk using a stick or crutches.

□ I am in bed most of the time and have to crawl to the toilet.

**Section 5: Sitting**

□ I can sit in any chair as long as I like.

□ I can sit in my favourite chair as long as I like.

□ Pain prevents me from sitting for more than 1 hour.

□ Pain prevents me from sitting for more than half an hour.

□ Pain prevents me from sitting for more than 10 minutes.

□ Pain prevents me from sitting at all.

**Section 6: Standing**

□ I can stand as long as I want without extra pain.

□ I can stand as long as I want but it gives me extra pain.

□ Pain prevents me from standing for more than 1 hour.

□ Pain prevents me from standing for more than half an hour.

□ Pain prevents me from standing for more than 10 minutes.

□ Pain prevents me from standing at all.

**Section 7: Sleeping**

□ My sleep is never disturbed by pain.

□ My sleep is occasionally disturbed by pain.

□ Because of pain I have less than 6 hours' sleep.

□ Because of pain I have less than 4 hours' sleep.

□ Because of pain I have less than 2 hours' sleep.

□ Pain prevents me from sleeping at all.

**Section 8: Sex life (if applicable)**

□ My sex life is normal and causes no extra pain.

□ My sex life is normal but causes some extra pain.

□ My sex life is nearly normal but is very painful.

□ My sex life is severely restricted by pain.

□ My sex life is nearly absent because of pain.

□ Pain prevents any sex life at all.

**Section 9: Social Life**

□ My social life is normal and causes me no extra pain.

□ My social life is normal but increases the degree of pain.

□ Pain has no significant effect on my social life apart from limiting my more energetic interests, e.g., sport, etc.
□ Pain has restricted my social life and I do not go out as often.
□ Pain has restricted social life to my home.
□ I have no social life because of pain.

**Section 10: Traveling**
□ I can travel anywhere without pain.
□ I can travel anywhere but it gives extra pain.
□ Pain is bad but I manage journeys over 2 hours.
□ Pain restricts me to journeys of less than 1 hour.
□ Pain restricts me to short necessary journeys under 30 minutes.
□ Pain prevents me from traveling except to receive treatment.

## Appendix 2: Roland Morris Questionnaire

When your back hurts, you may find it difficult to do some things you normally do.

This list contains some sentences that people have used to describe themselves when they have back pain. When you read them, you may find that some stand out because they describe you today. As you read the list, think of yourself today. When you read a sentence that describes you today, put a tick against it. If the sentence does not describe you, then leave the space blank and go on to the next one. Remember, only tick the sentence if you are sure it describes you today.

1. I stay at home most of the time because of my back.
2. I change position frequently to try and get my back comfortable.
3. I walk more slowly than usual because of my back.
4. Because of my back I am not doing any of the jobs that I usually do around the house.
5. Because of my back, I use a handrail to get upstairs.
6. Because of my back, I lie down to rest more often.
7. Because of my back, I have to hold on to something to get out of an easy chair.
8. Because of my back, I try to get other people to do things for me.
9. I get dressed more slowly than usual because of my back.
10. I only stand up for short periods of time because of my back.
11. Because of my back, I try not to bend or kneel down.
12. I find it difficult to get out of a chair because of my back.
13. My back is painful almost all of the time.
14. I find it difficult to turn over in bed because of my back.
15. My appetite is not very good because of my back pain.
16. I have trouble putting on my socks (or stockings) because of the pain in my back.
17. I only walk short distances because of my back pain.
18. I sleep less well because of my back.
19. Because of my back pain, I get dressed with help from someone else.
20. I sit down for most of the day because of my back.
21. I avoid heavy jobs around the house because of my back.
22. Because of my back pain, I am more irritable and bad tempered with people than usual.
23. Because of my back, I go upstairs more slowly than usual.
24. I stay in bed most of the time because of my back.

## Appendix 3: Head-to-Head Quality Criteria Form

Created by Anastasia Newman

**Article Citation**:

________________________________________________________________________________

________________________________________________________________________________

________________________________________________________________________________

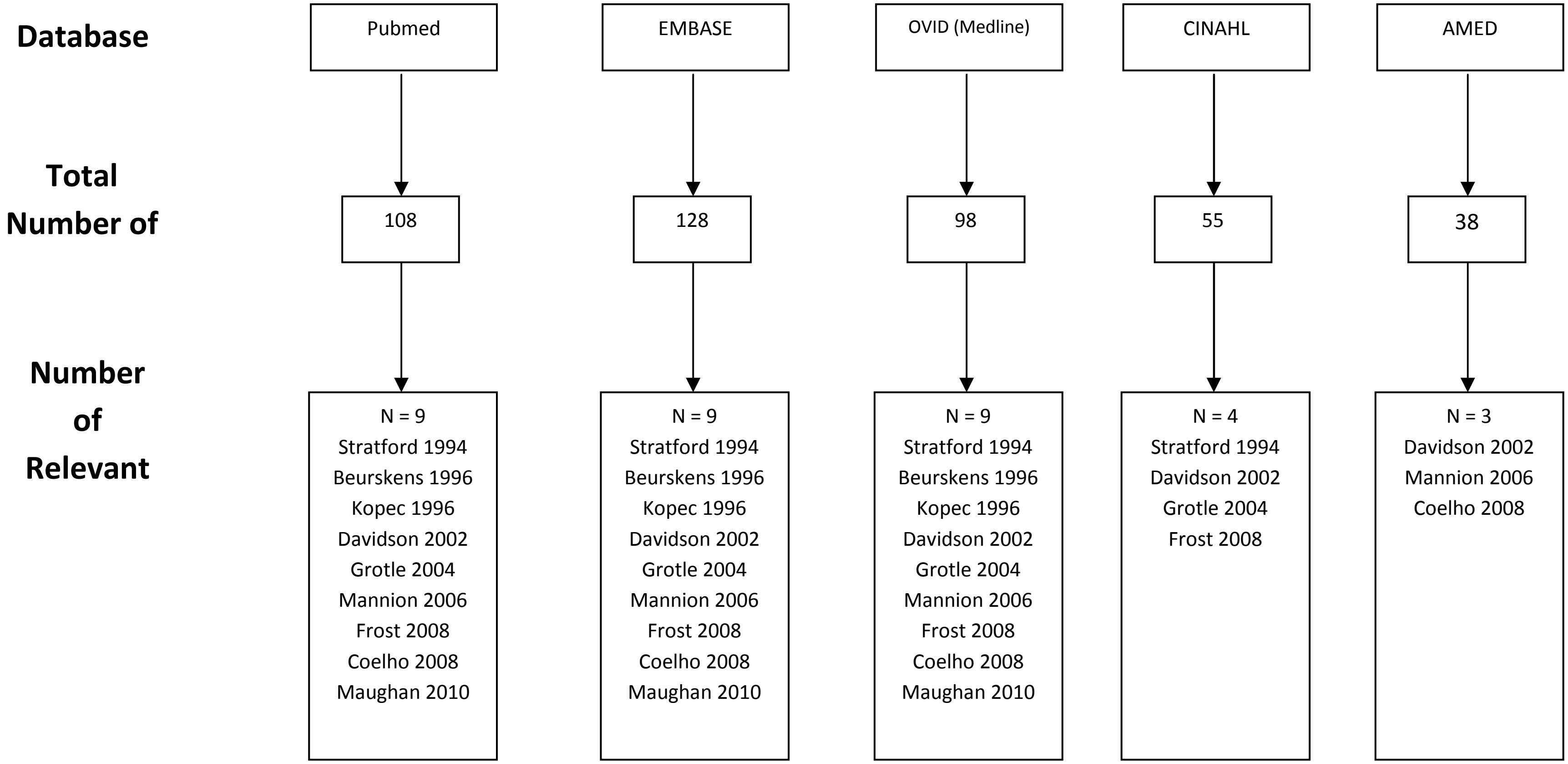
| Criteria | Yes | No | Can't Tell | N/A |
|---|---|---|---|---|
| **PURPOSE** | | | | |
| Was the purpose/research question clearly stated? | | | | |
| **METHODS** | | | | |
| **SAMPLE CHARACTERISTICS** | | | | |
| Were the eligibility criteria clearly stated? | | | | |
| Was the pain site noted? (eg, 12 to gluteal fold, radiating pattern) | | | | |
| **STUDY DESIGN** | | | | |
| Was the study design explicitly stated? | | | | |
| Was the setting of the study stated? | | | | |
| Was the country of study stated? | | | | |
| Were the measurement conditions similar for both measures? | | | | |
| Was the interval between assessments specified? | | | | |
| Was this interval between assessments justified? | | | | |
| Was the expectation of the sample's change characteristics stated? | | | | |
| Was a reference standard identified? | | | | |
| Was the reference standard independent of measures? | | | | |
| **MEASURE DESCRIPTION** | | | | |
| Was the language of the questionnaire stated? | | | | |
| **Language used:** | | | | |
| Was the version of the measure stated? | | | | |
| **Please state the versions used:** | | | | |
| Were any modifications made to the measures? | | | | |
| **Please list modifications:** | | | | |
| **SAMPLE SIZE** | | | | |
| Was a formal sample size calculation done? | | | | |
| What was the sample size? N = | | | | |
| **ANALYSIS** | | | | |
| Was the choice of analysis consistent with the samples expected change characteristics? | | | | |
| Was a formal comparison of the measures attempted? | | | | |
| Did the formal analysis account for dependent data? | | | | |
| **RESULTS** | | | | |
| Did the authors provide descriptive statistics of age? | | | | |

| Criteria | Yes | No | Can’t Tell | N/A |
|---|---|---|---|---|
| Were descriptive statistics provided concerning the duration of back pain prior to the study? | | | | |
| Was a gender distribution provided? | | | | |
| Were all co-morbidities reported (if applicable)? | | | | |
| Was the patients work distribution provided (if applicable)? | | | | |
| Was the distribution of pain pattern provided? | | | | |
| Was the proportion of unanswered/multiple response questions reported? | | | | |
| Did the authors comment on patient follow-up losses? | | | | |
| Were descriptive statistics of each measure provided for pre-scores? | | | | |
| Were descriptive statistics of each measure provided for post-scores? | | | | |
| Were descriptive statistics of each measure provided for change scores? | | | | |
| Were individual measure change statistics with p-value/confidence intervals provided (e.g.; SRM, ROC curve area, correlation coefficient)? | | | | |
| Were between measure comparison statistics with p-value/confidence intervals provided? | | | | |
| **CONCLUSION** | | | | |
| Were the authors’ conclusions consistent with the results? | | | | |

**ADDITIONAL COMMENTS:**

## Appendix 4: Flow Diagram

### Search Terms: “Roland” AND “Oswestry”



| Database | Pubmed | EMBASE | OVID (Medline) | CINAHL | AMED |
|---|---|---|---|---|---|
| Total Number of | 108 | 128 | 98 | 55 | 38 |
| Number of Relevant | N = 9<br>Stratford 1994<br>Beurskens 1996<br>Kopec 1996<br>Davidson 2002<br>Grotle 2004<br>Mannion 2006<br>Frost 2008<br>Coelho 2008<br>Maughan 2010 | N = 9<br>Stratford 1994<br>Beurskens 1996<br>Kopec 1996<br>Davidson 2002<br>Grotle 2004<br>Mannion 2006<br>Frost 2008<br>Coelho 2008<br>Maughan 2010 | N = 9<br>Stratford 1994<br>Beurskens 1996<br>Kopec 1996<br>Davidson 2002<br>Grotle 2004<br>Mannion 2006<br>Frost 2008<br>Coelho 2008<br>Maughan 2010 | N = 4<br>Stratford 1994<br>Davidson 2002<br>Grotle 2004<br>Frost 2008 | N = 3<br>Davidson 2002<br>Mannion 2006<br>Coelho 2008 |

## Appendix 5: Summary of Quality Criteria Results

| Criteria | Stratford | Kopec | Beurskens | Davidson | Grotle (a) | Grotle (b) | Mannion | Coelho | Frost | Maughan |
|---|---|---|---|---|---|---|---|---|---|---|
| **Purpose** | | | | | | | | | | |
| Was the purpose/research question clearly stated? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| **Sample Characteristics** | | | | | | | | | | |
| Were eligibility criteria clearly stated? | Y | Y | N | Y | Y | Y | N | Y | Y | Y |
| Was the pain site noted? | N | Y | Y | Y | Y | Y | N | N | N | N |
| **Study Design** | | | | | | | | | | |
| Was the study design explicitly stated? | Y | N | Y | Y | Y | Y | N | N | Y | Y |
| Was the setting of the study stated? | Y | Y | N | Y | Y | Y | Y | Y | Y | Y |
| Was the country of study stated? | N | Y | N | N | Y | Y | N | Y | Y | Y |
| Were the measurement conditions similar for both measures? | Y | CT | CT | CT | CT | CT | CT | CT | CT | CT |
| Was the interval between assessments specified? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Was the interval between assessments justified? | N | N | N | Y | N | N | N | N | N | N |
| Was the expectation of the sample's change characteristics stated? | N | Y | N | N | Y | Y | N | N | N | N |
| Was a reference standard identified? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Was the reference standard independent of measures? | N | N | N | N | Y | N | N | N | N | N |
| **Measure Description** | | | | | | | | | | |
| Was the language of the questionnaire stated? | N | Y | N | N | Y | Y | Y | Y | N | Y |

| **Criteria** | **Stratford** | **Kopec** | **Beurskens** | **Davidson** | **Grotle (a)** | **Grotle (b)** | **Mannion** | **Coelho** | **Frost** | **Maughan** |
|---|---|---|---|---|---|---|---|---|---|---|
| Language used | E | E/F | D | E | NW | NW | G | BP | E | E |
| Was the version of the measure stated? | Y | N | Y | Y | Y | Y | Y | Y | Y | Y |
| Version used (ie: RMQ-Original; RMQ-Modified; ODI v.1.0; ODI v.2.0; ODI v.2.1) | RMQ orig<br>ODI 1.0 | RMQ orig<br>ODI 1.0 | RMQ orig<br>ODI 1.0 | RMQ orig<br>ODI 2.0 | RMQ mod<br>ODI 2.0 | RMQ mod<br>ODI 2.0 | RMQ mod<br>ODI 2.1 | RMQ mod<br>ODI 1.0 | RMQ mod<br>ODI 2.0 | RMQ mod<br>ODI 2.0 |
| Were any modifications made to the measures? | N | CT | N | N | N | N | N | N | N | N |
| **Sample Size** | | | | | | | | | | |
| Was a formal sample size calculation done? | N | N | N | N | N | N | N | N | N | N |
| Sample size* | 88 | 178 | 76 | 101 | 100 | 100 | 57 | 30 | 201 | 48 |
| **Analysis** | | | | | | | | | | |
| Was the choice of analysis consistent with the samples expected change characteristics? | Y | CT | Y | Y | Y | Y | Y | Y | Y | Y |
| Was a formal comparison of the measures attempted? | Y | N | N | Y | N | N | N | N | N | N |
| Did the formal analysis account for dependent data? | N | N | N | Y | N | Y | N | N | N | N |
| **Results** | | | | | | | | | | |
| Did the authors provide descriptive statistics of age? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Were descriptive statistics provided concerning the duration of back pain prior to the study? | Y | Y | Y | Y | Y | Y | N | Y | N | Y |
| Was a gender distribution provided? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Were all co-morbidities reported? | N | N | N | N | N | N | N | N | N | N |

| Criteria | Stratford | Kopec | Beurskens | Davidson | Grotle (a) | Grotle (b) | Mannion | Coelho | Frost | Maughan |
|---|---|---|---|---|---|---|---|---|---|---|
| Were the patients' work distributions provided? | Y | Y | N | Y | Y | Y | N | N | N | Y |
| Was the distribution of pain pattern provided? | N | Y | Y | Y | Y | Y | N | N | N | N |
| Was the proportion of unanswered/multiple response questions reported? | Y | Y | Y | N | Y | Y | N | N | N | N |
| Did the authors comment on patient follow-up losses? | N | N | Y | Y | N | N | N | N | N | N |
| Were descriptive statistics of each measure provided for pre-scores? | Y | N | Y | Y | Y | Y | Y | Y | Y | Y |
| Were descriptive statistics of each measure provided for post-scores? | Y | N | Y | Y | N | N | Y | Y | Y | Y |
| Were descriptive statistics of each measure provided for change scores? | Y | N | Y | Y | Y | Y | Y | Y | Y | Y |
| Were individual measure change statistics with p-value/confidence intervals provided? | Y | N | N | Y | Y | Y | Y | Y | Y | N |
| Were between measure comparison statistics with p-value/confidence intervals provided? | Y | N | N | Y | Y | Y | N | N | N | Y |
| **Conclusion** | | | | | | | | | | |
| Were the authors' conclusions consistent with the results? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |

Y = Yes; N = No; CT = Can't tell; E = English; NW= Norwegian; D = Dutch; F = French; BP = Brazilian-Portuguese; G = German

* Sample sizes analyzed for this thesis/review

Grotle (a) = Expected Clinical Course as reference standard

Grotle (b) = Global Change Index as reference standard

## Glossary of Terms

Absolute Reliability: The extent to which a score varies on repeated measurements.  It is quantified by the standard error of the measurement.

Analysis of Variance (ANOVA): A family of statistical tests used to analyze differences between two or more groups.  It is based on partitioning the sum of square into that attributable to between-group differences and that attributable to within-group differences.

Ceiling Effect: a measurement phenomenon in which an instrument cannot register gains in scores for the participants of interest.

Construct Validity: The extent to which a test behaves in accordance with hypotheses concerning how it should behave.

Content Validity: The extent to which a test provides a comprehensive representation of the concept of interest.

Correlation Coefficient: Mathematical expression of the degree of relationship between two or more variables.

Criterion Validity: The extent to which a test agrees with a gold standard's assessment of the construct of interest.

Effect Size: Measures the ability of a health measurement scale to detect a signal (improvement) among the noise (variance) of a patient sample.  It is calculated by dividing the mean change score of the patient population by the standard deviation of the baseline scores.

Floor Effect: a measurement phenomenon in which an instrument cannot register greater declines in scores for the participants of interest.

Internal Consistency: One form of reliability.  It is the extent to which items on a test are homogenous.

Interrater Reliability: The extent to which multiple raters provide consistent ratings on a specific measure.

Intrarater Reliability: The consistency with which one rater assigns scores to a single set of responses on two or more occasions.

Minimal Clinically Important Difference (MCID): The smallest change that represents an important difference to either the client or the management of the client. It is expressed in the same units as the original measurement.

Minimal Detectable Change (MDC): This represents an estimate of the smallest change that can be detected for a client, expressed in the same units as the original measurement. It is based on the standard error of the measurement and a confidence level is usually included.

Outcome Measure: A measurement tool used to document change in one or more constructs over time.

Psychometric Properties: The reliability and validity, including responsiveness and sensitivity to change, of a measurement tool.

Receiver Operating Curve: Method of graphing test data to determine cut-off points that balance sensitivity and specificity.

Relative Reliability: Exists when individual measurements within a group maintain their position within the group on repeated measurements. It is quantified by correlation coefficients.

Reliability: The consistency of a measure; the extent to which measurements are repeatable. It is defined as the variance among subjects divided by the total variance.

Responsiveness: The ability of a measure to assess clinically important change over time. Often used interchangeably with sensitivity to change.

Self-Report Outcome Measure: A questionnaire completed by the client or someone acting on behalf of the client.

Sensitivity: The percentage of individuals with a particular diagnosis who are correctly identified by a test as having the condition of interest.

Sensitivity to Change: The capacity of a measure to assess change over time.

Specificity: The percentage of individuals without a particular diagnosis who are correctly identified by a test as not having the condition of interest.

Standard Error of Measurement: Represents the standard deviation of measurement errors; one way of measuring absolute reliability.  It is expressed in the same units as the original measurement.

Standardized Response Mean: The mean change score divided by the standard deviation of the change scores.

Systematic Review: A form of literature review that requires a documented search strategy and explicit inclusion and exclusion criteria for studies reviewed, reducing author bias toward or against particular methods or outcomes.

Test-retest Reliability: The extent to which multiple applications of a test provide consistent results.

Validity: The extent to which a measure assesses what it is intended to measure.