ASSESSING PHYSICAL FUNCTIONING IN LOW BACK PAIN

ASSESSING PHYSICAL FUNCTION IN LOW BACK PAIN

By MAYSA ALNATTAH, BSc PT, MSc MS, PgD PH

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Master's Degree of Science in Rehabilitation Sciences

MASTER OF SCIENCE (2021) in Rehabilitation Science McMaster University, Hamilton, Ontario, Canada

TITLE: Assessing physical function in low back pain

AUTHOR: Maysa Alnattah PT, BSc, MSc

SUPERVISOR: Dr. Luciana G. Macedo

SUPERVISORY COMMITTEE: Dr. Ayse Kuspinar
Dr. Marla Beauchamp

NUMBER OF PAGES: 165

Lay Abstract

Low Back Pain (LBP) care costs the Canadian health care system millions of dollars every year. Most clinicians and researchers use self-report questionnaires filled out by their patients to assess physical function. However, performance measures where patients perform tasks while being observed are also recommended to assess physical function. Performance-based measures can be used alone or in combination with self-report measures. To select the most appropriate performance measures, we need to know how good and trustworthy these measures are. Therefore, the purpose of this study was to collect all possible performance measures that were developed or used to assess physical function in LBP patients; then summarized the available evidence on their psychometric properties (reliability, validity and responsiveness).

We searched five scientific databases and found 47 studies that evaluated 115 performance measures. Most included studies were of low quality and evaluated different tests or test properties. We found that most measures were not reliable, accurate or were sensitive to change. Therefore, clinicians and researchers need caution when selecting and interpreting results of these performance measures when evaluating physical function in LBP.

Abstract

Physical function has been identified as a core outcome to be assessed in low back pain (LBP). However, all recommended physical function measures are Patient-Reported Outcome Measures (PROMs). Performance-Based Measures (PBMs) are important measures that are practical and are prone to fewer biases. Two systematic reviews provided evidence on the psychometric properties of PBMs but were not comprehensive. Therefore, the purpose of this study was to identify PBMs developed for or used to assess physical function in LBP and to review studies evaluating the psychometric properties of these PBMs systematically.

The first manuscript of the thesis was the systematic review protocol developed using the COSMIN (COnsensus-based Standards for the selection of health status Measurement INstruments) manual 2018. The protocol was also registered on PROSPERO (CRD42020147968). The protocol also outlined the use of the COMINS Risk of Bias (COSMIN-ROB) checklist 2018; standard priory hypotheses and criterions developed to evaluate the results of each psychometric property; as well as a GRADE criterion (Grading of Recommendations, Assessment, Development and Evaluations) to assess the level of evidence. Two reviewers independently screened, evaluated, and extracted data.

The second manuscript was the systematic review written in the format of a journal for future submission. Our database search identified 47 studies assessing 115 PBMs. In general, findings included five different LBP diagnoses (e.g., non-specific LBP) and different LBP durations (e.g., acute, chronic). The level of evidence of each PBM or psychometric property mainly were generated from single studies. A high risk of bias assessed by the COSMIN-ROB checklist was found for most of the included studies. Overall, the included studies' results often did not meet our priory hypotheses for good psychometric properties. Hence, most PBMs' psychometric properties were found to have a low level of evidence. There was not a single PBM that demonstrated a good level of evidence for all properties. In conclusion, significant heterogeneity was found between studies leading to a limited level of evidence. PBMs need to be used with great caution. High-quality studies that investigate PBMs' psychometric properties are needed.

Acknowledgements

First, I would like to give my most gratitude and appreciations to the most important person who have made this work possible and helped me succeed, my incredible mentor and supervisor Dr. Luciana Macedo. Without her reassurance, patience, and great support, I wouldn’t have succeeded. She was a mentor and a friend who took up with my difficult times and health issues and showed great encouragement and direction to make sure I come to a successful end in this long journey.

Second, I would like to extend my thanks and regards to my committee members, Dr. Ayse Kuspinar and Dr. Marla Beauchamp, who took the time to review my work and upgrade my work by their valuable comments and guidance. I appreciate your constant feedback and answering my emails whenever I needed the help.

And, for the most important people in my life, my husband Dr. Richard A Gosselin and my best friend Maryam Alqadery, who without them, I wouldn’t have the strength and the ability to go through this journey. You have always stood by my side in the good and bad days. Through these past 3 years, you have overwhelmed me with your endless support, patience, encouragement and love. I thank you for the rest of my life for being in my life.

Finally, I would also like to thank my friend and colleagues Dr. Nora Bakaa, Cassandra D'Amore, and Shannon Killip who helped with data screening, evaluation and extraction. This work would not have been possible without your assistance.

Declaration of Academic Achievement

For all the manuscripts, Maysa Alnattah developed the research questions, designed the studies, collected and analyzed the data, and wrote the initial drafts.

Chapter 3:

Dr. Luciana Macedo helped refine the research questions, data collection, extraction, and analysis.

Dr. Luciana Macedo, Dr. Ayse Kuspinar, and Dr. Marla Beauchamp helped develop the research questions, research methods, and reviewing/editing and providing feedback for the manuscripts.

Table of Contents

List of Tables

Chapter 1: Introduction to Epidemiology, Risk Factors, Prognosis, and Management of Low Back Pain

Chapter 3: A Systematic Review of the psychometric Properties of Performance-Based Measures to assess Physical Function in Low Back Pain patients.

List of Figures

Chapter 1: Introduction to Epidemiology, Risk Factors, Prognosis, and Management of Low Back Pain

Chapter 3: A Systematic Review of the psychometric Properties of Performance-Based Measures to assess Physical Function in Low Back Pain patients.

List of Appendices

Chapter Two: A Protocol for a Systematic Review of Measurement Characteristics of Physical Performance-Based Measures for individuals with Low Back Pain

List of Abbreviations

OR – Odd Ratio

AUC – Area Under the Curve

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

COSMIN – COnsensus-based Standards for the selection of health Measurement INstruments

COSMIN-ROB – COSMIN-Risk of Bias checklist 2018

WHO- ICF – World Health Organization's International Classification of Functioning, Disability and Health Model

COMET – Core Outcome Measures in Effectiveness Trials

OMERACT – Outcome Measures in Rheumatology

ODI – Oswestry Disability Index

RMDQ – Roland Morris Disability Questionnaire

NRS – Numeric Rating Scale

SF12 – Short Form Health Survey 12

PROMs – Patient Reported Outcome Measures

PROMIS-GH-10 – 10-item PROMIS Global Health

HRQoL – Health-Related Quality of Life

PBMs – Performance-Based Measures

ICC – Intraclass correlation coefficient

GRADE – Grading of Recommendations, Assessment, Development and Evaluations

k – Kappa

SDC – Smallest Detectable Change

LoA – Limit of Agreement

MIC – Minimal Important Change

GRS – Global Rating Scale

Chapter One: Introduction

1.1 Low Back Pain

Low Back Pain (LBP) is defined as pain or discomfort typically located in the lower back region (between the lower rib margins and the gluteal folds).[1 2] This pain can be accompanied by loss of spine range of motion, stiffness, uni- or bi-lateral leg pain and other associated neurological symptoms in the lower limbs (e.g. radiculopathy).[1 2] LBP can be associated with significant loss of physical function and disability, leading to poor health-related quality of life.[3 4] The ability to return to full work, participate in life situations, and emotional and mental health can all be compromised due to LBP.[3 4]

LBP can be classified into Specific-LBP, Non-Specific LBP and Serious Pathology related-LBP.[5] Specific-LBP includes conditions for which specific pathoanatomical aetiologies for symptoms can be identified, such as canal stenosis and degenerative disc disease.[5 6] Non-Specific LBP is LBP with no determined pathoanatomical causes, and it is the most common form of LBP (85% of cases).[6 7] Serious pathologies related to LBP have a prevalence of less than 1 % and include cancer, fractures, and infections.[5]

LBP can also be classified according to its duration into acute (less than 6 weeks), sub-acute (between 6 and 12 weeks) and chronic (12 weeks or more).[5] However, there is significant criticism around this classification, given the contemporary view that LBP is a long-term health condition with episodes of recurrence, remission, and flares.[8 9] It means individuals with LBP might experience fluctuating or persistent pain, making it difficult to categorize new exacerbations into acute or chronic.[8 9] Hence, LBP's simplistic classification into the three aforementioned categories may not capture the complete scope of LBP trajectories over time.[8]

1.2 Epidemiology and Burden

According to the latest Global Burden of Disease Study 1990-2017, LBP has been the leading cause of Years Lived with Disability (YLDs) for nearly three decades.[10] LBP was responsible for ≈ 65 million YLDs in 2017 for both sexes combined globally, representing a 17.5 % increase since 2007 and the highest attribution among the three leading causes of YLD's counts.[10] LBP's burden also represents significant challenges to the health care system and the economy at global and national levels.[10] In Canada, the total annual LBP-related estimate of medical costs (Direct-Costs) is $6-12 billion, not including societal costs associated with disability payment and loss of worker productivity (Indirect-Costs).[11] In Australia, the indirect costs of LBP are estimated to be sixfold of the direct costs.[12]

In 2017, approximately 577 million people were affected by LBP globally.[10] A recent systematic review that included data from Canada, the United States of America (USA), Sweden, Belgium, Finland, Israel, and the Netherlands, estimated LBP prevalence and incidence from studies that used electronic medical records.[13] The mean point prevalence estimates for LBP ranged from 1.4% to 20 % (50-80 % of adult life), with higher prevalence observed among industry workers (aerospace, defense industry, space technology and telecommunication).[13] In the same review, the incidence of LBP ranged from 20% to 28%.[13] Similarly, LBP incidence was higher in industrial workers,[13] indicating that occupation load may be a risk factor for LBP.[13]

1.3 LBP Risk Factors

LBP is a multifactorial condition known to have multiple risk factors.[14] [15] Risk factors can be categorized into intrinsic (within individual factors) and extrinsic (environmental factors).[14] [15] Intrinsic factors include sex, age, Body Mass Index, and poor general health (e.g.

presence of comorbidities).[14 15] Extrinsic factors include occupation,[14 15] lifestyle,[16] or social factors.[17-19]

A systematic review that summarized longitudinal cohort studies identified age, sex, height, BMI, smoking, physical activity level, history of back pain, job satisfaction, and structural imaging as risk factors for LBP (see table 1 for a summary of potential risk factors).[6 15 20 21] However, most studies had low quality of evidence and conflicting results across risk factors.[6 15 17 20 22 23] The only consistently identified LBP risk factor was having *a history of back pain,* [6 15 20 21] with LBP recurrence rates at one-year follow-up ranging from 24% to 54%.[24-28]

Table 1: Summary of previously identified risk factors for low back pain

Type of factors	**Risk Factors**
Occupational Factors	- Type of jobs (e.g., heavy physical strain, frequent lifting, postural stress, and vibration).[14 15] - Long working hours.[19] - Psychosocial factors: poor attitude towards the employer, low job satisfaction, poor worker-supervisor interaction, low monotony at work, job control and security, absence of social support and work-family balance, hostile work environment, and no decision authority.[14-16]
Psychological Factors [23]	- Anxiety and Depression. - Catastrophizing. - Kinesophobia (fear of movement). - Somatization (the expression of distress as physical symptoms or their persistence).
Demographic/lifestyle factors	- Gender (conflicting results between male and female).[22] - Older age. - Low education. - single marital status. - high BMI/obesity. - Smoking. - Alcohol consumption. - Poor general health. - Low physical fitness.[14-16]
History of Back Pain	- History of LBP in the past is the strongest and most consistently identified risk factor for having another LBP episode or transition to chronicity in the future.[6 15 20 21]

1.4 Natural history and Prognosis

As aforementioned, LBP is a long-term condition with episodes of recurrence, remission, and flares.[8 9] Usually, 90% of individuals with new acute LBP episodes recover within the first 6 to 12 weeks.[4 5 29-31] Although most studies evaluating prognostic factors in LBP are of poor quality or have weak methodologies,[4 30] some putative prognostic factors have been linked to short recovery time and rapid return to work.[4 32-34] Prognostic factors for acute LBP can include physical fitness (exercising or playing sports), nature of LBP (sudden onset with no previous history of LBP), occupational (high job satisfaction and have been working pre-injury) and personal factors (high education level, self-referred to doctor).[4 30 32-34]

Although a large number of individuals recover from acute-LBP, approximately one third will experience a recurrence within one year.[21 34] In general, one out of five acute LBP patients will develop chronic LBP (persistent LBP) in Canada.[21 34] There is no robust evidence for the risk to transition to chronicity; however, in an attempt to create a screening questionnaire to predict transition to persistent LBP, Traeger AC et al. 2016 developed the Predicting the Inception of Chronic Pain (PICKUP) tool. The five prognostic factors, screened by PICKUP, were found to increase the chance of chronicity such as disability compensation (Odd Ratio (OR): 1.65); leg pain (OR: 1.56); pain intensity (OR: 1.23); depression (OR: 1.06); and perceived risk (OR: 1.14). [35] Perceived risk reflects the individuals' judgments of the risk for LBP persistence.[34-36] However, the tool was found to have inadequate predictive validity for (Area Under the Curve (AUC) = 0.66 [95% CI 0.63 to 0.69]) identifying those at risk for chronicity.[35] According to COSMIN-Risk of Bias (COSMIN-ROB) checklist 2018, an AUC of 0.70 is considered acceptable and the PICKUP tool did not meet that value; hence, results from the PICKUP tool should be interpreted with caution.

1.5 LBP Management

For the past 20 years, there has been an increase in the number of clinical practice guidelines for LBP management published worldwide.[37] LBP management can be conservative or non-conservative depending on LBP's etiology, degree of pain intensity, and loss of function.[7] [38-40] Conservative treatments are usually the first line of care, and other interventions such as surgery are offered when conservative care has failed, or symptoms and activity limitation are severed.[7] [38-40] Conservative treatments include patient education and self-care (e.g., advice to stay active), exercise therapy (e.g., strengthening, core exercise), manual therapy, cognitive behavioural therapy, massage, acupuncture, mindfulness and pharmacological therapies (e.g., paracetamol, non-steroidal anti-inflammatory).[7] [38-40] Non-conservative treatments usually include surgery.[7] [38-40]

The selection of LBP treatment is usually based on the duration of symptoms and etiology.[40] For acute LBP, treatment aims to reduce pain, prevent transitioning to persistent LBP, and reduce and prevent associated disability.[40] Clinical practice guidelines often suggest that the first line of care is advice to stay active, including return to usual daily activities such as work.[40] However, early supervised exercise therapy is advised for patients who are declining or who have risk factors for transitioning to chronicity.[40] For chronic LBP, treatments are focused on improving quality of life and preventing further decline in function.[40] The most recommended intervention for chronic LBP is exercise therapy.[40] There is no specific type of exercise that has shown to be superior to another; therefore, exercise programs should be tailored according to patients' preferences, capabilities and needs, and functionally focused.[40] In addition, multidisciplinary rehabilitation, including cognitive behavioural therapy, should be offered to persons at risk for poor prognosis.[40 41] Finally, there is no strong evidence to support the use of pharmacological medication such as paracetamol and opioids.[40]

As mentioned previously, surgery is only recommended for persons for whom conservative management has failed or for those with severe symptoms and activity limitation.[40] However, surgery often does not lead to full recovery, and approximately one-third of patients will have long-term pain and disability.[40]

1.6 Outcome Measures in Rehabilitation

An outcome is defined as a construct of interest to be measured, and it is sometimes represented by a latent variable that cannot be directly observed (e.g., physical function).[42] An outcome measure is a tool used to measure the construct of interest.[42] Outcome measures are often used in research and clinical practice for three primary purposes.[43 44] First, they are used to evaluate changes in patients' health over time following a specific treatment.[43 44] Second, they can be used for diagnostic purposes to discriminate between different groups, such as to classify patients into different treatments.[43 44] Third, they assist in predicting patients' future health status.[43 44]

Many aspects should be considered before selecting and using an outcome measure, including acceptable psychometric properties (reliability, validity and responsiveness).[44 45] COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) is an initiative of an international multidisciplinary team of researchers who are experts in developing and evaluating outcome measures.[45 46] COSMIN's primary goal is to provide clinicians with reliable and valid instruments, and researchers with transparent methodologies to properly evaluate these instruments.[45 46] To prevent confusion resulting from using different definitions of psychometric terminologies, COSMIN provides a consensus-based taxonomy of measurement properties for reliability, validity and responsiveness.[45 47]

According to COSMIN, reliability is defined as "the degree to which the measurement is free from measurement error."[45 47] Reliability reflects on the extent to which individuals' scores stay the same at repeated measurements under stable conditions. It measures the variation in repeated measurements.[45 47] Hence, it provides researchers and clinicians with an indication of confidence in the measurements used to quantify a construct of interest (e.g. pain, walking, balance).[45 47] Measurements of variation can include test-retest (patients provide the same scores under several conditions); inter-rater (two or more raters provide the same scores of the same group of patients); intra-rater (same rater provides the same scores under several conditions); or internal consistency (multi-items test measure different aspects of the same construct).[45 47] Depending on the type of reliability being assessed, variations can result from the patients being assessed; clinicians conducting the assessment; circumstances at the time of measurements; or instruments used to measure the construct.[45 47]

Validity is another critical psychometric property that measures "the degree to which an instrument measures the construct(s) it aims to measure."[45 47] There are three major types of validity: content validity, criterion validity, and construct validity.[45 47] Content validity is the extent to which an instrument's content appears to reflect the construct of interest concerning relevance and comprehensiveness.[45 47] Criterion validity is the degree of agreement between a measurement instrument's scores and the scores of a gold standard instrument for the construct of interest.[45 47] When there is no gold standard instrument for the construct of interest, construct validity is measured instead.[45 47] Construct validity depends on existing knowledge and hypothesis about the construct and reflects on an instrument's ability to provide the expected scores.[45 47]

In addition to validity and reliability, outcome measures have to be responsive and able to detect change in patients' health status over time in the construct of interest.[45 47] This refers to responsiveness, which COSMIN defines as "the ability of an instrument to detect change over time in the construct to be measured."[45 47] For example, if patients' pain levels change, then the scores on a measurement instrument assessing pain should change accordingly.[45 47]

Besides demonstrating good psychometric properties, other qualities could be considered when developing or selecting an outcome measurement instrument: [43]

- be convenient and comfortable for patients;
- be applicable across different populations and in different contexts;
- have consistent and defined protocols to follow and interpret;
- be able to achieve its purpose (evaluation/assessment, diagnosis, classification, prognosis/declining, prediction);
- reflect the affected health domain that is of interest; and
- possess comparative data to norms with others with similar conditions.[43]

1.7 Conceptual Model

The International Classification of Functioning, Disability and Health (ICF) is the most internationally accepted model, introduced by the WHO (World Health Organization) in 2001, to provide a global description of the disablement process across different countries and cultures. The WHO-ICF's disablement process provides a structured model to simplify the process of describing, classifying, and measuring Function and Health.[46 48] The WHO-ICF model starts with the person's *Health Condition*.[48] It then describes the disablement process of this health condition using three main domains: *Body Function and Body Structure*; *Activity;* and *Participation*.[48] It also considers the interaction of these domains with *Environmental Factors* and *Personal Factors* that impact the individual's ability to function in everyday life (see figure 1).[48]

Body Function and Body Structure domain focuses on the anatomical (structure) and physiological (function) parts of the body system (e.g. sensory, motor, neuromusculoskeletal functions).[48] It describes abnormalities at cellular or organ levels that might lead to impairments.[48] *Activity* and *Participation* domains focus on the individual and social dimensions, respectively.[48] *Activity* is a domain that reflects human daily life tasks, such as tasks related to mobility, movement and self-care.[48] Impairments at the *Body Function and Body Structure* domain cause loss of function in the affected body part, which then affect *Activity*.[48] A loss in *Activity* is referred to as disability.[48] On the other hand, *Participation* is a domain that deals with human communication with the outer world and the interaction within his/her society, such as work and employment, or personal or social relationships.[48] Loss in persons' *Activity,* which leads to disability, does not necessarily affect individuals' participation in society, meaning that a disabled person can still have a job; hence, participating in his society. *Environmental* and *Personal* factors can either trigger, reduce or exacerbate impairments and disabilities.[48] Accordingly, loss in function or health occurs at three different levels (biological, individual, and social) and is affected by other contextual factors. The WHO-ICF model provides a scientific ground to standardize reporting of outcomes and assess interventions' efficacy and effectiveness across clinical practice and research.

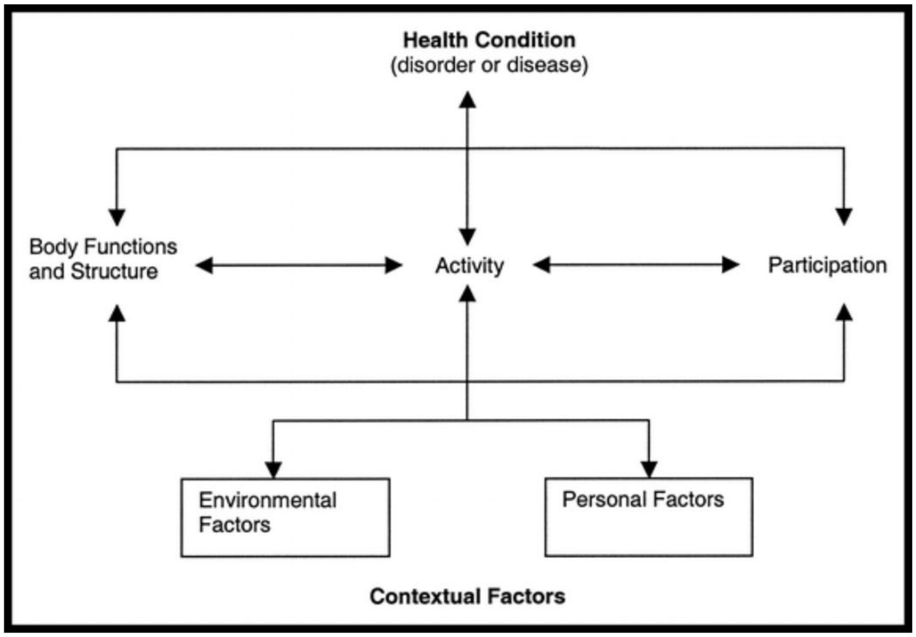


Figure 1: World Health Organization's International Classification of Functioning, Disability, and Health Model (WHO-ICF).

1.8 LBP Core Outcome Set

Over the past few decades, the number of available outcome measures has increased dramatically. Consequently, it is difficult to select the most appropriate instrument for use in clinical practice and research.[49 50] In addition, this can lead to inconsistency in reporting outcomes between clinical trials, which makes it difficult to compare results and conduct meta-analyses in future systematic reviews. Therefore, understanding the different components of the WHO-ICF's disablement process allows researchers to identify important outcomes when evaluating a health condition.[48]

Different international committees have used the WHO-ICF as a ground model to build a consensus-based set of outcomes to comprehensively and consistently evaluate health conditions within clinical trials and research.[43 51] This set of outcomes is usually referred to as a Core Outcome Set.[51] A Core Outcome Set allows standardization across different disciplines and contexts and facilitates effective interdisciplinary communication and efficient multidisciplinary team care and research.[43 51]

An international committee, involving researchers, clinicians, and patient representatives from different countries, was formed to develop a Core Outcome Set for LBP.[52] The Core Outcome Measures in Effectiveness Trials (COMET) and Outcome Measures in Rheumatology (OMERACT) meet regularly to discuss, standardize, and generate consensus-based core outcome sets to be used in LBP-research and clinical practice based on the WHO-ICF model.[52-54] Using a robust Delphi methodology, four outcome domains were identified as being essential for measurement in LBP trials: physical function; pain intensity; health-related quality of life; and number of deaths.[53 54]

Following the initial Delphi study, a systematic review and a second Delphi study were conducted to identify instruments to be used within each Core Outcome Set domain. The decision to include instruments was based on their psychometric properties. All included instruments were PROMs and the author's justification was because of the PROMs feasibility as well as because they are the most frequently used and recommended instruments in the LBP literature.[53 54] These instruments included the Oswestry Disability Index (ODI 2.1 a) and Roland Morris Disability Questionnaire (RMDQ-24) for assessing physical function; the Numeric Rating Scale (NRS) with a 1-week recall period for assessing pain intensity; the Short Form Health Survey 12 (SF12) and 10-item PROMIS Global Health (PROMIS-GH-10) for assessing health-related quality of life; and a simple statement for reporting any death occurred in a clinical trial.[53] Some of the psychometric properties of the Patient Reported Outcome Measures (PROMs) mentioned above are summarized in table 2 A-C.

Table 2-A: Psychometric properties of the PROs to assess physical function in LBP (ODI & RMDQ)

Measurement properties		Result Rating	Quality of evidence
Measurement properties for ODI			
Content validity [55]	**Relevance**	Inconsistent Results	Very Low
	Comprehensiveness	Unsatisfactory Results	Very Low
	Comprehensibility	Satisfactory Results	Very Low
Structural validity [55]		Inconsistent Results	Moderate
Internal consistency [56]		Unknown	Unknown
Reliability [56]		Satisfactory Results	Moderate
Measurement error [56]		Satisfactory Results	Moderate
Construct validity [56]		Inconsistent Results	Conflicting
Responsiveness [56]		Inconsistent Results	Conflicting
Measurement properties for RMDQ			
Content validity[55]	**Relevance**	Satisfactory Results	Very Low
	Comprehensiveness	Unsatisfactory Results	High
	Comprehensibility	Satisfactory Results	High
Structural validity[55]		Unsatisfactory Results	High
Internal consistency[56]		Unknown	Unknown
Reliability[56]		Inconsistent Results	Conflicting
Measurement error[56]		Unsatisfactory Results	Moderate
Construct validity[56]		Satisfactory Results	Moderate
Responsiveness[56]		Inconsistent Results	Conflicting

PF, Physical Function; ODI, Oswestry Disability Index version; RMDQ, Roland Morris Disability Questionnaire.

Table 2-B: Psychometric properties of the Pain Intensity Scale (NRS) in LBP

Measurement properties for NRS [55]		Result Rating	Quality of evidence
Content validity	**Relevance**	Inconsistent Results	Low

	Comprehensiveness	Inconsistent Results	Low
	Comprehensibility	Satisfactory Results	Very Low
Structural validity		NA	NA
Internal consistency		NA	NA
Test-retest reliability		Inconsistent Results	Low
Measurement error		Unsatisfactory Results	High
Construct validity		Inconsistent Results	Very Low
Responsiveness		Inconsistent Results	Moderate

NRS, Numeric Rating Scale; NA, Not Applicable

Table 2-C: Psychometric properties of Health-Related Quality of Life PROs in LBP (SF-12 and PROMIS-GH-10)

Measurement properties		**Result Rating**	**Quality of evidence**
Measurement properties for SF-12			
Content validity [55]	**Relevance**	Satisfactory Results	Very Low
	Comprehensiveness	Satisfactory Results	Very Low
	Comprehensibility	Satisfactory Results	Very Low
Structural validity [57]		NA	NA
Internal consistency [57]		NA	NA
Reliability [57]		NA	NA
Measurement error [57]		NA	NA
Construct validity [55]	**PCS**	Inconsistent Results	Low
	MCS	Inconsistent Results	Low
Responsiveness [55]	**PCS**	Inconsistent Results	Very Low
	MCS	Unsatisfactory Results	Low
Measurement properties for PROMIS-GH-10			
Content validity [55]	**Relevance**	Satisfactory Results	Very Low
	Comprehensiveness	Satisfactory Results	Very Low
	Comprehensibility	Satisfactory Results	Very Low
Structural validity [57]		NA	NA
Internal consistency [57]		NA	NA
Reliability [57]		NA	NA
Measurement error [57]		NA	NA
Construct validity [57]		NA	NA
Responsiveness [57]		NA	NA

HRQoL, Health-Related Quality of Life; SF-12 (PCS and MCS), Short Form Health Survey 12 (PCS= Physical Component Score, and MCS = Mental Component Score); PROMIS-GH-10: Patient-Reported Outcomes Measurement Information System Global Health 10-items; NA, Not available due to lack of sufficient number of studies.

1.9 Physical Function in LBP Research

Physical function is a multi-dimensional and highly complex construct,[58] and the meaning of physical function varies greatly between individuals.[58] As stated earlier, "dysfunction", according to the WHO-ICF model, can occur at three different domains; *Body structure and function*, *Activity*, and *Participation*.[48 58] Physical function is described interchangeably by the two domains; *Activity* and *Participation*.[48 58] Physical function represents both human physical performance and individuals' role in society, and the interconnection between these two constructs.[48 58] In real life, it is very challenging to separate human performance from participation.[48 58] However, for the current thesis context, we simplify the definition of physical function to follow as much as possible the WHO-ICF *Activity* domain.[48 58] Therefore, physical function is defined as "any restriction or lack of ability to perform a task or an activity in the manner considered normal for a person."[48 58]

Given the significant worldwide disability impact of LBP,[40 48 58] measuring physical function in this patient group is essential in providing data on the impact of LBP on individuals' lives for both research and clinical practice.[40 48 58] Therefore, this thesis's objectives will be to explore existing outcome measures of physical function in LBP.

1.9.1 Patient-Reported Outcome Measures to Assess Physical Function

The classifications (types) of outcome measures greatly vary depending on what they are measuring, how they are reported, the dimensionality of the measured construct, and what theoretical framework is used.[45] PROMs are defined as "a measurement of any aspect of a patient's health that comes directly from the patient without interpretation of the patient's responses by a physician or anyone else."[59] PROMs are the most routinely used outcome measures in LBP studies.[45]

PROMs do not involve an examiner/observer;[45] therefore, they can be collected using paper format or electronically (e.g., phone applications, e-mail).[60] They are efficient, economical, and time saving.[60] Given its often easy application, they may be applied multiple times to track changes in patients' outcomes.[60] Nevertheless, PROMs suffer from many limitations and biases. They are subjective measures that can suffer from social desirability bias, recall bias and are often influenced by individual's contextual and psychosocial factors such as fear-avoidance and catastrophizing.[61] [62] PROMs' responses may not reflect the actual patients' physical ability; but the individual's perception of their abilities.[63-65] Further, some PROMs can be long and complex and may be difficult to collect in patients with language, communication and educational barriers.[60] Psychometric properties of core PROMs for LBP are reported in section 1.2.

1.9.2 Performance-Based Measures to Assess Physical Function

Performance-Based Measures (PBMs) are measurement that are performed by patients and observed by clinicians. Sometimes these instruments are described as "Objective".[45] However, objective tests are tests in which clinicians and patients do not influence the outcomes, such as X-rays or blood tests.[45] Assessing physical function cannot be entirely considered objective, and a level of subjectivity is often present.[45] Most physical PBMs have the potential to be influenced by assessors (e.g., instructions to patients, measurement error) and patients (e.g., Hawthorne effects, polite patient bias).[45] Consequently, different assessors, contexts and instructions might influence patients' motivations when performing these measurements.[45] Nonetheless, PBMs overcome some of the limitations observed with PROMs, such as the influence of personal factors (e.g., patients' education levels or language skills) and possible discrepancies between capabilities and perceived capabilities.[63 66 67] PBMs are measurements with predefined criteria in which a patient is asked to perform a standardized task that assessors can observe and quantify outcomes (e.g. time, distance, repetitions).[68] Therefore, PBMs provide unique information of a patients' physical performance that is complementary to a PROM.[68]

Psychometric Properties

Two recent systematic reviews summarized the psychometric properties of PBMs used to assess physical function in the LBP population.[69 70] The first systematic review, published in 2018, was focused on reliability only.[69] The review included 20 studies that identified 38 outcome measures. However, not all 38 measurements were PBMs or evaluated physical function.[69] For example, the review included muscle strength and muscle endurance tests.[69] The PBMs that were most commonly evaluated in the review were the 50-Ft Walking Test and The Sit-To-Stand Test.[69] Other identified PBMs were the 5-Minute Walk Test, Time-Up-And-Go Test, Shuttle Walk Test, Stair Climbing, and Progressive Isoinertial Lifting Evaluation.[69] Four of the PBMs (5-Minute Walking Test, 50-Ft Walking Test, Shuttle Walk Test, and Sit-To-Stand Test) had good test-retest reliability with an ICC ranging from 0.76 to 0.99.[69] However, most included studies were of low methodological quality of "Fair" (11 studies) to "Poor" (9 studies) according to COSMIN-ROB (Rick of Bias) checklist 2012.[69] For intra-rater reliability, only 50-Ft Walking Test, Time-Up-And-Go Test, and 30-sec Chair Stand Test had high ICC values ranging from 0.94 to 0.99.[69] Good inter-rater reliability (ICC= 0.98-0.99) was found for 50-Ft Walking Test, Sit-To-Stand Test, Time-Up-And-Go Test, and Walking Ability; and good agreement with kappa ranging from 0.76-1.0 for Sock Test, Pick-Up Test, and Roll-Up Test.[69] Nonetheless, these results of inter- and intra-reliability were reported by only one study each.[69]

Another systematic review was published in 2019.[70] It was a more comprehensive review, summarizing reliability, validity and responsiveness properties and included PBMs of physical function aligned with the ICF definition of *Activity* domain definition.[70] There were 25 studies included in the review with 18 PBMs.[70] The results of this review were consistent with the review published in 2018 on reliability.[69] [70] Twelve studies investigated convergent validity.[70] Six PBMs had good convergent validity (75% of the results met the hypotheses), including the 50-ft walk task, 5-minute walk task, modified lift test, Progressive Isoinertial Lifting Evaluation, 5-repetition sit-to-stand task, and Timed "Up & Go" task.[70] In total, there were only seven studies that assessed responsiveness; of these, only three PBMs met the pre-specified hypothesis (i.e., 1-minute stair-climbing task, shuttle walking test, and 5-repetition sit-to-stand test); but most of the studies evaluated had poor methodological quality (COSMIN checklist 2012).[70] In general, the methodological quality of the included studies was low, and only a few PBMs met the predefined criteria (e.g., hypothesis for construct validity) for validity and responsiveness.[70]

In general, these two systematic reviews were of good quality.[69] [70] They followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines in their reporting and used the COSMIN-checklist 2012 when evaluating the methodological quality of studies.[69] [70] In the first review, many of the included measures did not reflect physical function and rather focused on endurance or joint movements (e.g., Extensor endurance test, Squatting, repeated trunk rotation);[69] and in the second review, battery tests and functional capacity tests were excluded.[70] In addition, many of the included PBMs were excluded from data synthesis due to the studies' poor grade on COSMIN-ROB checklist 2012.[70] Other limitations also existed, such as narrow search terms, and excluding non-English studies and grey literature.[69 70] In addition to the aforementioned limitations, both reviews used the 2012 edition of the COSMIN-ROB checklist and not the updated 2018 edition, which is less conservative than the 2012 version.[69 70]

1.10 Thesis purpose and objectives

This thesis aimed to conduct a systematic review of PBMs of physical function for people with low back pain. This thesis will be more comprehensive and improve on the quality of previously published reviews by: 1) using a comprehensive definition of physical function based on the WHO-ICF model's definition of *Activity* with consideration of the overlap between "*Activity*" and "*Participation*" aspect of physical function; 2) using the updated COSMIN-ROB 2018 checklist to assess the risk of bias of included studies; 3) using a robust method for interpreting results (results rating) with the pre-specification of the hypothesis for each psychometric property, particularly for construct validity and responsiveness; and 4) having no restrictions on language.

The objectives of this thesis were to:

- Identify physical Performance-Based Measures developed or used to assess physical function in low back pain patients.
- Synthesize the available evidence on the psychometric properties of the identified physical Performance-Based Measures using COSMIN-ROB 2018 checklist.

Chapter 2 includes a comprehensive protocol for the review that was developed a priori and registered with PROSPERO (CRD42020147968). Chapter 3 includes the full systematic review written in the submission format for the Journal of Orthopedic and Sports Physical Therapy. Chapter 4 includes a summary, strengths and limitations and future recommendations.

1.11 References

1. Hoy D, Brooks P, Blyth F, et al. The epidemiology of low back pain. *Best practice & research Clinical rheumatology* 2010;24(6):769-81.

2. Freburger JK, Holmes GM, Agans RP, et al. The rising prevalence of chronic low back pain. *Archives of internal medicine* 2009;169(3):251-58.

3. Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: Summary and general recommendations. *Spine* 2000;25(24):3100-03.

4. Pengel LH, Herbert RD, Maher CG, et al. Acute low back pain: systematic review of its prognosis. *Bmj* 2003;327(7410):323.

5. Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. *The Lancet* 2018;391(10137):2356-67.

6. Riihimäki H. Low-back pain, its origin and risk indicators. *Scandinavian journal of work, environment & health* 1991:81-90.

7. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet* 2017;389(10070):736-47.

8. Kongsted A, Kent P, Axen I, et al. What have we learned from ten years of trajectory research in low back pain? *BMC musculoskeletal disorders* 2016;17(1):220.

9. Macedo LG, Maher CG, Latimer J, et al. Nature and determinants of the course of chronic low back pain over a 12-month period: a cluster analysis. *Phys Ther* 2014;94(2):210-21. doi: 10.2522/ptj.20120416 [published Online First: 2013/09/28]

10. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* 2018;392(10159):1789-858.

11. CANADA BAJ. Low Back Pain 2014 [Available from: http://boneandjointcanada.com/low-back-pain/.

12. Hoy D, March L, Brooks P, et al. Measuring the global burden of low back pain. *Best practice & research Clinical rheumatology* 2010;24(2):155-65.

13. Fatoye F, Gebrye T, Odeyemi I. Real-world incidence and prevalence of low back pain using routinely collected data. *Rheumatology international* 2019;39(4):619-26.

14. Manchikanti L. Epidemiology of low back pain. *Pain physician* 2000;3(2):167-92.

15. Manchikanti L, Singh V, Falco FJ, et al. Epidemiology of low back pain in adults. *Neuromodulation: Technology at the Neural Interface* 2014;17:3-10.
16. Chou Y-C, Shih C-C, Lin J-G, et al. Low back pain associated with sociodemographic factors, lifestyle and osteoporosis: a population-based study. *Journal of rehabilitation medicine* 2013;45(1):76-80.
17. Buruck G, Tomaschek A, Wendsche J, et al. Psychosocial areas of worklife and chronic low back pain: a systematic review and meta-analysis. *BMC musculoskeletal disorders* 2019;20(1):480.
18. Melloh M, Elfering A, Stanton TR, et al. Who is likely to develop persistent low back pain? A longitudinal analysis of prognostic occupational factors. *Work* 2013;46(3):297-311.
19. Yang H, Haldeman S, Lu M-L, et al. Low back pain prevalence and related workplace psychosocial risk factors: a study using data from the 2010 National Health Interview Survey. *Journal of manipulative and physiological therapeutics* 2016;39(7):459-72.
20. Øiestad BE, Hilde G, Tveter AT, et al. Risk factors for episodes of back pain in emerging adults. A systematic review. *European Journal of Pain* 2020;24(1):19-38.
21. Da Silva T, Mills K, Brown BT, et al. Risk of recurrence of low back pain: a systematic review. *journal of orthopaedic & sports physical therapy* 2017;47(5):305-13.
22. Bento TPF, dos Santos Genebra CV, Maciel NM, et al. Low back pain and some associated factors: is there any difference between genders? *Brazilian Journal of Physical Therapy* 2020;24(1):79-87.
23. Beynon AM, Hebert JJ, Hodgetts CJ, et al. Chronic physical illnesses, mental health disorders, and psychological features as potential risk factors for back pain from childhood to young adulthood: a systematic review with meta-analysis. *European Spine Journal* 2020:1-17.
24. Biering-Sørensen F. A prospective study of low back pain in a general population. I. Occurrence, recurrence and aetiology. *Scandinavian journal of rehabilitation medicine* 1983;15(2):71-79.
25. Carey TS, Garrett JM, Jackman A, et al. Recurrence and care seeking after acute back pain: results of a long-term follow-up study. *Medical care* 1999:157-64.
26. Cassidy JD, Côté P, Carroll LJ, et al. Incidence and course of low back pain episodes in the general population. *Spine* 2005;30(24):2817-23.

27. Stanton TR, Henschke N, Maher CG, et al. After an episode of acute low back pain, recurrence is unpredictable and not as common as previously thought. *Spine* 2008;33(26):2923-28.

28. Stevenson JM, Weber CL, Smith JT, et al. A longitudinal study of the development of low back pain in an industrial population. *Spine (03622436)* 2001;26(12):1370-77.

29. Croft PR, Dunn KM, Raspe H. Course and prognosis of back pain in primary care: the epidemiological perspective: LWW, 2006.

30. Costa LdCM, Maher CG, Hancock MJ, et al. The prognosis of acute and persistent low-back pain: a meta-analysis. *Cmaj* 2012;184(11):E613-E24.

31. Bakker EW, Verhagen AP, Lucas C, et al. Spinal mechanical load: a predictor of persistent low back pain? A prospective cohort study. *European Spine Journal* 2007;16(7):933-41.

32. Coste J, Lefrançois G, Guillemin F, et al. Prognosis and quality of life in patients with acute low back pain: insights from a comprehensive inception cohort study. *Arthritis Care & Research* 2004;51(2):168-76.

33. Faber E, Burdorf A, Bierma-Zeinstra SM, et al. Determinants for improvement in different back pain measures and their influence on the duration of sickness absence. *Spine* 2006;31(13):1477-83.

34. Henschke N, Maher CG, Refshauge KM, et al. Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. *Bmj* 2008;337:a171.

35. Traeger AC, Henschke N, Hübscher M, et al. Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low back pain. *PLoS medicine* 2016;13(5)

36. Hayden J, Dunn K, Van der Windt D, et al. What is the prognosis of back pain? *Best Practice & Research Clinical Rheumatology* 2010;24(2):167-79.

37. Dagenais S, Tricco AC, Haldeman S. Synthesis of recommendations for the assessment and management of low back pain from recent clinical practice guidelines. *The Spine Journal* 2010;10(6):514-29.

38. Borenstein D. Epidemiology, etiology, diagnostic evaluation, and treatment of low back pain. *Current Opinion in Orthopaedics* 1999;10(2):131-36.

39. Balagué F, Mannion AF, Pellisé F, et al. Non-specific low back pain. *The lancet* 2012;379(9814):482-91.

40. Foster NE, Anema JR, Cherkin D, et al. Prevention and treatment of low back pain: evidence, challenges, and promising directions. *The Lancet* 2018;391(10137):2368-83.

41. Morso L, Kongsted A, Hestbaek L, et al. The prognostic ability of the STarT Back Tool was affected by episode duration. *European Spine Journal* 2016;25(3):936-44.

42. Boers M, Kirwan JR, Wells G, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *Journal of clinical epidemiology* 2014;67(7):745-53.

43. Law MC, MacDermid J. Evidence-based rehabilitation: A guide to practice: Slack Incorporated 2008.

44. Riddle D, Stratford P. Is this change real?: Interpreting patient outcomes in physical therapy: FA Davis 2013.

45. de Vet HC, Terwee CB, Mokkink LB, et al. Measurement in medicine: a practical guide: Cambridge University Press 2011.

46. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018;78:1.

47. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of clinical epidemiology* 2010;63(7):737-45.

48. Organization WH. International Classification of Functioning, Disability and Health (ICF) 2003 [updated 2 March 2018. Available from: https://www.who.int/classifications/icf/en/.

49. Thonnard JL, Penta M. Functional assessment in physiotherapy. A literature review. *Europa Medicophysica* 2007;43(4):525-41.

50. Taylor AM, Phillips K, Patel KV, et al. Assessment of physical function and participation in chronic pain clinical trials: IMMPACT/OMERACT recommendations. *Pain* 2016;157(9):1836-50. doi: 10.1097/j.pain.0000000000000577

51. Chiarotto A, Ostelo RW, Turk DC, et al. Core outcome sets for research and clinical practice. *Brazilian journal of physical therapy* 2017;21(2):77-84.

52. Chiarotto A, Terwee CB, Deyo RA, et al. A core outcome set for clinical trials on non-specific low back pain: study protocol for the development of a core domain set. *Trials* 2014;15(1):511.

53. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain* 2018;159(3):481-95.

54. Chiarotto A, Deyo RA, Terwee CB, et al. Core outcome domains for clinical trials in non-specific low back pain. *European Spine Journal* 2015;24(6):1127-42.

55. Chiarotto A, Terwee CB, Ostelo RW. A systematic review on the content validity of questionnaires to measure physical functioning in patients with low back pain. *Quality of Life Research* 2016;25 (1 Supplement 1):142-43.

56. Chiarotto A, Maxwell LJ, Terwee CB, et al. Roland-Morris Disability Questionnaire and Oswestry Disability Index: Which Has Better Measurement Properties for Measuring Physical Functioning in Nonspecific Low Back Pain? Systematic Review and Meta-Analysis. *Physical Therapy* 2016;96(10):1620-37.

57. Chiarotto A, Terwee CB, Kamper SJ, et al. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: A systematic review. *Journal of Clinical Epidemiology* 2018;102:23-37. doi: 10.1016/j.jclinepi.2018.05.006

58. O'Sullivan SB, Schmitz TJ, Fulk G. Physical rehabilitation: FA Davis 2019.

59. Mayo NE. ISOQOL Dictionary of Quality of Life and Health Outcomes Measurement. First edition ed: International Society for Quality of Life Research (ISOQOL) 2015.

60. Philpot LM, Barnes SA, Brown RM, et al. Barriers and benefits to the use of patient-reported outcome measures in routine clinical care: a qualitative study. *American Journal of Medical Quality* 2018;33(4):359-64.

61. Stone AA, Shiffman S, Schwartz JE, et al. Patient compliance with paper and electronic diaries. *Controlled clinical trials* 2003;24(2):182-99.

62. Buer N, Linton SJ. Fear-avoidance beliefs and catastrophizing: occurrence and risk factor in back pain and ADL in the general population. *Pain* 2002;99(3):485-91.

63. Simmonds MJ, Olson SL, Jones S, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine* 1998;23(22):2412-21.

64. Lee CE, Simmonds MJ, Novy DM, et al. Self-reports and clinician-measured physical function among patients with low back pain: A comparison. *Archives of Physical Medicine and Rehabilitation* 2001;82(2):227-31. doi: 10.1053/apmr.2001.18214

65. Conway J, Tomkins CC, Haig AJ. Walking assessment in people with lumbar spinal stenosis: capacity, performance, and self-report measures. *Spine Journal* 2011;11(9):816-23. doi: 10.1016/j.spinee.2010.10.019

66. Gautschi OP, Smoll NR, Corniola MV, et al. Validity and reliability of a measurement of objective functional impairment in lumbar degenerative disc disease: The Timed Up and Go (TUG) test. *Neurosurgery* 2016;79(2):270-78.

67. Caporaso F, Pulkovski N, Sprott H, et al. How well do observed functional limitations explain the variance in Roland Morris scores in patients with chronic non-specific low back pain undergoing physiotherapy? *European Spine Journal* 2012;21:S187-S95. doi: 10.1007/s00586-012-2255-6

68. Harding VR, de C Williams AC, Richardson PH, et al. The development of a battery of measures for assessing physical functioning of chronic pain patients. *Pain* 1994;58(3):367-75.

69. Denteneer L, Van Daele U, Truijen S, et al. Reliability of physical functioning tests in patients with low back pain: a systematic review. *Spine Journal* 2018;18(1):190-207.

70. Jakobsson M, Gutke A, Mokkink LB, et al. Level of Evidence for Reliability, Validity, and Responsiveness of Physical Capacity Tasks Designed to Assess Functioning in Patients With Low Back Pain: A Systematic Review Using the COSMIN Standards. *Physical Therapy* 2019;99(4):457-77. doi: 10.1093/ptj/pzy159

Chapter Two: A Protocol for a Systematic Review of Measurement Characteristics of Physical Performance-Based Measures for individuals with Low Back Pain

Title: A protocol of a systematic review of psychometric properties of physical performance-based measures to assess physical functioning in individuals with low back pain

Authors:
Maysa Alnattah, PT BSc, MS MSc, PH PgD, School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Luciana G. Macedo, PT Ph.D., School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Ayse Kuspinar, PT Ph.D., School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Marla Beauchamp, PT Ph.D., School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Corresponding author:
Maysa Alnattah, alnattam@mcmaster.ca, room 308, IAHS, McMaster University, Hamilton, ON, Canada

2.0 Abstract

Background: Physical function is primally assessed using Patient-reported outcome measures (PROMs) in low back pain (LBP). However, physical function should also be assessed using performance-based measures (PBMs). The purpose of this study will be to identify PBMs developed/used to assess physical function in LBP population and to systematically review studies evaluating the psychometric properties of these PBMs.

Methods: Five databases will be searched (MEDLINE, EMBASE, AMED, CINAHL, and SPORTDiscus) using four search term-domains (LBP, performance tests/measurers, physical function, and psychometric properties). Studies will be included if they recruited individuals with LBP (with no serious pathology), used PBMs to assess physical function, and investigated any psychometric properties of these measurements. Two authors will complete study screening, evaluation, and data extraction. Data synthesis will be based on a pre-established criterion for results rating and the COnsensus-based Standards for the selection of health status Measurement INstruments) Risk of Bias score (COSMIN-ROB).

Discussion and Conclusion: This systematic review will enable researchers and clinicians to identify and select the most appropriate PBMs for assessing physical function in LBP.

Keywords: low back pain, psychometric property, physical function, performance-based measures.

2.1 Introduction

Physical function is one of the Core Outcome Sets recommended to be measured in Low Back Pain (LBP) trials.[1] Physical function was defined by the International Classification of Functioning, Disability and Health (ICF) as “any restriction or lack of ability to perform a task or an activity in the manner considered normal for a person”.[2 3] There are various outcome measures available to assess physical function, including Patient-Reported Outcome Measures (PROM) or physical Performance-Based measures (PBM).[4-8] PROMs are the most commonly used and recommended tools to measure physical function in LBP. Nevertheless, they suffer from many limitations such as recall bias,[9-15] and have a low level of evidence for psychometric properties (e.g., criterion validity).[16 17] Therefore, the use of both PROMs and PBMs to comprehensively assess physical function is highly recommended.[17]

To our knowledge, there have been two systematic reviews on the psychometric properties of PBMs used to assess physical function in LBP.[18 19] In general, they were well conducted; however, some important limitations were identified. Both used the old version of COSMIN-ROB checklist of 2012 (COnsensus-based Standards for the selection of health status Measurement Instruments-Risk of Bias) and therefore excluded high risk of bias studies, excluded battery tests and Functional Capacity Evaluation tests, excluded non-English language manuscripts, or were focused on only one psychometric property (i.e., reliability).[18 19] Hence, the purpose of the current systematic review protocol is to overcome some of these limitations and to conduct a comprehensive systematic review of PBMs used to assess physical function in LBP.

2.2 Methods

2.2.1 Study design:

The review followed the updated COSMIN systematic review methodology manual 2018 for conducting the current review, assessing the included studies' methodological quality, and data synthesis of level of evidence.[20]

Objectives:

- To identify PBMs that have been developed or used to assess physical function in LBP patients.
- To synthesize the available evidence on the psychometric properties of the identified PBMs.

Search:

A search will be conducted to identify: 1) studies that reported on the development or use of physical function PBMs for LBP patients, and 2) studies that evaluated the psychometric properties of physical function PBMs in LBP patients. Studies that have evaluated the psychometric properties of PBMs in individuals with LBP, even if the tool was not developed for LBP, will be included.

The following databases will be used: MEDLINE (OvidSP, 1946 to 19 May 2019); EMBASE (OvidSP, 1980 to 19 May 2019); AMED (OvidSP, 1985 to 19 May 2019); CINAHL (EBSCO, 1981 to 19 May 2019); and SPORTDiscus (EBSCO, 1800 to 19 May 2019). The search terms that will be used will span the following domains: low back pain, physical performance-based Measurements, physical function, and psychometric properties. Once we identify outcome measures from the previous search, we will also include a search for psychometric properties using the name of the outcome measure and psychometric property terms. See appendix (A) for search terms.

Hand searches of reference lists of similar systematic reviews and all included studies will be conducted. We will also conduct citation tracking of included studies using ISI Web of Science.

2.2.2 Eligibility Criteria:

Type of studies:

Studies that have developed or used at least one PBM of physical function in LBP. We will also include studies that have tested the psychometric properties of the PBMs identified. No restriction of language or publication date will be imposed. We will not impose limits to study designs.

Type of participants:

Studies that have included participants of age equal and above 18 years old and any sex/gender who have LBP will be considered. LBP is defined as pain or discomfort in the lower back region attributed to any known or unknown pathoanatomical cause.[21-24] Studies that have included participants with specific and non-specific LBP of any duration will be included. Studies on the development of a PBM of physical function that use mixed populations in relation to diagnosis (e.g., OA) will only be included if data for LBP patients can be presented separately. Authors of included studies that used mixed populations (e.g., specific and non-specific LBP) will be contacted to provide data for each sub-group; however, the study will be excluded if data cannot be provided.

Type of outcome measurement:

PBMs that are intended to measure physical function in LBP patients will be considered. Physical Performance-Based Measures are Measurements and tools that health professionals and researchers use to collect information about patients' current physical presentations.[25 26] These

Measurements are used to assess, measure and observe patients' actual performance on a set of functional tasks, including: 1) self-care skills, 2) transfers, 3) mobility, 4) movements and so on.[25 26] PBMs are distinct from PROMs, which are directly obtained or filled out by the patients themselves.[27] Examples of PROMs are questionnaires, such as the Roland-Morris Disability Questionnaire and the Oswestry Disability Index.[1]

Types of outcome:

Physical function

In this systematic review, Physical function is defined according to the International Classification Of Functioning Disability and Health (ICF), which falls under the umbrella terms of patients' *Activity* (with the consideration of overlapping with *Participation* domain).[28] PBMs of physical function will be considered as Measurements of *Activity* if they are used to determine patients' ability to execute a task or action in a safe and timely manner. In addition, to be eligible for inclusion in data synthesis, measurements should have a clear protocol with pre-specified measurement units (e.g., kg, seconds) or total scores. [28] Examples of PBMs are the Timed–Up and Go Test, Sit-to-Stand, and Tinetti Mobility Test. PROMs and Impairment-Based Measurements (e.g. Range of Motion, muscle performance) will be excluded.[29]

Psychometric properties

The three domains of the psychometric properties that will be extracted from the studies are reliability, validity and responsiveness as defined by COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN) checklist 2018.[20]

2.2.3 Screening:

The screening of titles and abstracts and potentially eligible studies' full texts will be conducted independently by two reviewers. An inclusion form developed before the start of the review will be used to assess eligibility (see appendix B). A third reviewer will resolve any disagreements.

2.2.4 Methodological Quality Assessment:

Psychometric properties will be identified and evaluated according to the COSMIN taxonomy and definitions. The updated COSMIN-ROB (Risk of Bias) checklist (2018) will be used to assess the risk of bias of the included studies. Two reviewers will independently evaluate and summarize the results of the included studies. A third reviewer will resolve any disagreements. The COSMIN checklist assesses standard requirements for study design and statistical methods for measurement studies of health-related instruments.

Reviewers extracting data will participate in informal training on using the screening tools and methodological quality assessment as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement recommendations before data extraction.

2.2.5 Data extraction:

Two reviewers will extract data independently into a data extraction form developed and pilot tested before the beginning of the study (see Appendix C-1 and C-2). Study characteristics will be summarized in a table format. The following will be extracted: author name, study design, language, objectives, inclusion and exclusion criteria, participant demographic information, sample size, physical PBMs used, and description of the content and scoring of the Measurements, outcome measures, and type of psychometric properties evaluated.

All results on reliability, validity, and responsiveness of the PBMs will be considered. Types of reliability can include test-retest, inter-rater, intra-rater, and internal consistency (if exist for PBMs). Types of validity can include criterion validity (predictive and concurrent), face and content validity (if existed for PBMs), and construct validity (convergent, known-group and discriminant). Responsiveness measures can include both criterion and construct approaches.

2.2.6 Data synthesis:

Data synthesis will be presented using a GRADE criteria (Grading of Recommendations, Assessment, Development and Evaluations) based on the COSMIN handbook. [20] The GRADE takes into consideration two levels of assessment: 1) COSMIN-ROB checklist assessment, and 2) result rating assessment. Result rating is a process of comparing psychometric properties identified to pre-determined cut-off points or hypotheses (see Appendix D-1.) Using both COSMIN ROB and results rating assessment, data synthesis will be conducted using the criteria in appendix D-2. For example, a PBM that had a grade "very good" by COSMIN-ROB and exceeded the pre-determined cut-off point for test-retest reliability (ICC ≥ 0.70) will have a strong level of evidence. However, if the PBM had an "inadequate" grade or did not exceed the cut-off point, the level of evidence will be poor.

2.3 Discussion and Conclusion

The current systematic review aims to identify PBMs that have been developed or used to assess physical function in LBP population. Therefore, the purpose of this review is to conduct a comprehensive review: 1) using the updated COSMIN-ROB checklist of 2018 to assess the risk of bias of included studies;[20] 2) determining a priori hypotheses for psychometric properties to rate the results of included PBMs;[20] and 3) including all possible PBMs that follow the WHO-ICF definition of physical function (e.g. functional capacity evaluation).

The strengths of this study include the use of the updated COSMIN-ROB checklist (2018) and the integration of these results in the assessment of level of evidence using GRADE's recommendation.[20] Further, the use of hypothesis testing provides an opportunity to compare identified results with a null hypothesis that will be established for each outcome measure and psychometric testing. Hypothesis testing will enhance the interpretation of the results.

Limitations of this study include the use of an informal method of translating non-English studies (e.g., google translate), which may lead to inadequate interpretations. Moreover, we propose using the COSMIN-ROB, which was constructed to evaluate PROMs; however, COSMIN suggests that the ROB checklist is sufficiently rigorous and applicable for assessing PBMs.

In conclusion, we will conduct a systematic review of the psychometric properties of available PBMs used to assess physical function in people with LBP. This study will include a robust methodology, different from the already available reviews. Similarly, this study will provide an in-depth and more detailed review of all PBMs with the inclusion of studies that might have been excluded from the previously published systematic reviews.[19] Finally, the findings that emerge from the current systematic review can subsequently form the theoretical and empirical basis for selecting outcome measures to be used in clinical practice and research for assessing physical function.

2.4 Reference

1. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain* 2018;159(3):481-95.
2. Organization WH. International Classification of Functioning, Disability and Health (ICF) 2003 [updated 2 March 2018. Available from: https://www.who.int/classifications/icf/en/.
3. O'Sullivan SB, Schmitz TJ, Fulk G. Physical rehabilitation: FA Davis 2019.
4. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine* 2011;36(21 SUPPL.):S54-S68.
5. Thonnard JL, Penta M. Functional assessment in physiotherapy. A literature review. *Europa Medicophysica* 2007;43(4):525-41.
6. Taylor AM, Phillips K, Patel KV, et al. Assessment of physical function and participation in chronic pain clinical trials: IMMPACT/OMERACT recommendations. *Pain* 2016;157(9):1836-50.
7. Carey TS, Mielenz TJ. Measuring outcomes in back care. *Spine* 2007;32(11 SUPPL.):S9-S14.
8. Turk DC, Melzack R. Trends and future directions in human pain assessment. 2001
9. Stone AA, Shiffman S, Schwartz JE, et al. Patient compliance with paper and electronic diaries. *Controlled clinical trials* 2003;24(2):182-99.
10. Buer N, Linton SJ. Fear-avoidance beliefs and catastrophizing: occurrence and risk factor in back pain and ADL in the general population. *Pain* 2002;99(3):485-91.
11. Heneweer H, Picavet HSJ, Staes F, et al. Physical fitness, rather than self-reported physical activities, is more strongly associated with low back pain: evidence from a working population. *European Spine Journal* 2012;21(7):1265-72.

12. Anderson B, Lygren H, Magnussen LH, et al. What Functional Aspects Explain Patients' Impression of Change after Rehabilitation for Long-lasting Low Back Pain? *Physiotherapy Research International* 2013;18(3):167-77.

13. Guildford BJ, Jacobs CM, Daly-Eichenhardt A, et al. Assessing physical functioning on pain management programmes: the unique contribution of directly assessed physical performance measures and their relationship to self-reports. *British journal of pain* 2017;11(1):46-57.

14. Tucker JM, Welk GJ, Beyler NK. Physical activity in US adults: compliance with the physical activity guidelines for Americans. *American journal of preventive medicine* 2011;40(4):454-61.

15. Mcloughlin MJ, Colbert LH, Stegner AJ, et al. Are women with fibromyalgia less physically active than healthy women? *Medicine and science in sports and exercise* 2011;43(5):905.

16. Chiarotto A. Patient-Reported Outcome Measures: Best Is the Enemy of Good (But What if Good Is Not Good Enough?). *The Journal of orthopaedic and sports physical therapy* 2019;49(2):39-42.

17. Chiarotto A, Ostelo RW, Boers M, et al. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. *Journal of Clinical Epidemiology* 2018;95:73-93.

18. Denteneer L, Van Daele U, Truijen S, et al. Reliability of physical functioning tests in patients with low back pain: a systematic review. *Spine Journal* 2018;18(1):190-207.

19. Jakobsson M, Gutke A, Mokkink LB, et al. Level of Evidence for Reliability, Validity, and Responsiveness of Physical Capacity Tasks Designed to Assess Functioning in Patients

With Low Back Pain: A Systematic Review Using the COSMIN Standards. *Physical Therapy* 2019;99(4):457-77. doi: 10.1093/ptj/pzy159

20. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018;78:1.
21. Savigny P, Watson P, Underwood M, et al. Early management of persistent non-specific low back pain: summary of NICE guidance. *BMJ* 2009;338:b1805.
22. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet* 2017;389(10070):736-47.
23. Hoy D, Brooks P, Blyth F, et al. The epidemiology of low back pain. *Best practice & research Clinical rheumatology* 2010;24(6):769-81.
24. Freburger JK, Holmes GM, Agans RP, et al. The rising prevalence of chronic low back pain. *Archives of internal medicine* 2009;169(3):251-58.
25. de Vet HCW, Terwee CB, Mokkink LB, et al. Measurement in Medicine: A Practical Guide. Cambridge: Cambridge University Press 2011.
26. Erickson M, Erickson ML, McKnight R, et al. Physical therapy documentation: from examination to outcome: Slack Incorporated 2008.
27. de Vet HC, Terwee CB, Mokkink LB, et al. Measurement in medicine: a practical guide: Cambridge University Press 2011.
28. Reiman MP, Manske RC. The assessment of function: How is it measured? A clinical perspective. *The Journal of manual & manipulative therapy* 2011;19(2):91-9. doi: 10.1179/106698111x12973307659546 [published Online First: 2012/05/02]
29. Reiman MP, Manske RC. The assessment of function: How is it measured? A clinical perspective. *Journal of Manual & Manipulative Therapy* 2011;19(2):91-99.

2.6 Appendixes

2.6.1 Appendix A: Database Search Terms and strategy

OVID MEDLINE:

1. exp Back Pain/
2. back pain.mp.
3. backache*.mp.
4. exp Low Back Pain/
5. low back pain.mp.
6. exp Lumbar Vertebrae/
7. exp Spondylolisthesis/
8. ((lumb$ or back) adj pain).ti,ab.
9. exp Sacrococcygeal Region/
10. exp Sciatica/
11. sciatic*.mp.
12. back-ache.mp.
13. exp Intervertebral Disc Degeneration/
14. lumbago.mp.
15. exp Intervertebral Disc Displacement/
16. dorsalgia.mp.
17. exp Spinal Diseases/
18. exp Spondylosis/
19. exp Spondylitis, Ankylosing/

20. exp Spondylolisthesis/
21. exp Spondylitis/
22. spondy*.mp.
23. exp Back Injuries/
24. back injur*.mp.
25. exp "Activities of Daily Living"/
26. activity of daily living.mp.
27. physical function*.mp.
28. limitation of activity*.mp.
29. daily living activity.mp.
30. exp Self Care/
31. exp Physical Functional Performance/
32. exp "Physical and Rehabilitation Medicine"/
33. physical.mp.
34. performance.mp.
35. abilit*.mp.
36. exp "Recovery of Function"/
37. function.mp.
38. exp Movement/
39. movement.mp.
40. exp Executive Function/
41. function* task*.mp.
42. exp Mobility Limitation/

43. mobility.mp.

44. function* status.mp.

45. exp Health Status/

46. function* abilit*.mp.

47. ADL.mp.

48. exp Walking/

49. exp Gait/

50. exp Physical Fitness/

51. performance based test*.mp.

52. performance-based test*.mp.

53. exp Psychomotor Performance/

54. performance test*.mp.

55. exp Stair Climbing/

56. objective measure*.mp.

57. exp Mobility Limitation/

58. limitation.mp.

59. standing.mp.

60. sitting.mp.

61. exp Community Participation/

62. exp Patient Participation/

63. physical task*.mp.

64. exp Patient Transfer/

65. exp PSYCHOMETRICS/

66. exp “Reproducibility of Results”/
67. exp VALIDATION STUDIES/
68. valid*.mp.
69. reliab*.mp.
70. exp “Sensitivity and Specificity”/
71. reproducib*.mp.
72. repeatability.mp.
73. responsiveness.mp.
74. sensitiv*.mp.
75. specificity.mp.
76. psychometr*.mp.
77. exp Spinal Stenosis/
78. spinal stenosis.mp.
79. or/1-24
80. 77 or 78 or 79
81. or/65-76
82. 34 or 35 or 37 or 39 or 43 or 58
83. 33 and 82
84. 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 36 or 37 or 38 or 40 or 41 or 42 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57 or 59 or 60 or 61 or 62 or 63 or 64
85. 83 or 84
86. 80 and 81 and 85

87. limit 86 to humans

OVID EMBASE:

1. exp Back Pain/
2. back pain.mp.
3. backache*.mp.
4. exp Low Back Pain/
5. low back pain.mp.
6. exp Lumbar Vertebrae/
7. exp Spondylolisthesis/
8. ((lumb$ or back) adj pain).ti,ab.
9. exp Sacrococcygeal Region/
10. exp Sciatica/
11. sciatic*.mp.
12. back-ache.mp.
13. exp Intervertebral Disc Degeneration/
14. lumbago.mp.
15. exp Intervertebral Disc Displacement/
16. dorsalgia.mp.
17. exp Spinal Diseases/
18. exp Spondylosis/
19. exp Spondylitis, Ankylosing/
20. exp Spondylolisthesis/
21. exp Spondylitis/

22. spondy*.mp.

23. exp “Activities of Daily Living”/

24. activity of daily living.mp.

25. physical function*.mp.

26. limitation of activity*.mp.

27. daily living activity.mp.

28. exp Self Care/

29. exp “Physical and Rehabilitation Medicine”/

30. physical.mp.

31. performance.mp.

32. abilit*.mp.

33. exp “Recovery of Function”/

34. function.mp.

35. exp Movement/

36. movement.mp.

37. function* task*.mp.

38. mobility.mp.

39. function* status.mp.

40. exp Health Status/

41. function* abilit*.mp.

42. ADL.mp.

43. exp Walking/

44. exp Gait/

45. exp Physical Fitness/

46. performance based test*.mp.

47. exp Psychomotor Performance/

48. performance test*.mp.

49. exp Stair Climbing/

50. objective measure*.mp.

51. exp Mobility Limitation/

52. limitation.mp.

53. standing.mp.

54. sitting.mp.

55. physical task*.mp.

56. exp Patient Transfer/

57. exp PSYCHOMETRICS/

58. exp "Reproducibility of Results"/

59. exp VALIDATION STUDIES/

60. valid*.mp.

61. reliab*.mp.

62. exp "Sensitivity and Specificity"/

63. reproducib*.mp.

64. repeatability.mp.

65. responsiveness.mp.

66. sensitiv*.mp.

67. specificity.mp.

68. psychometr*.mp.
69. 57 or 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68
70. 31 or 32 or 34 or 36 or 38 or 52
71. 30 and 70
72. exp vertebral canal stenosis/
73. 23 or 24 or 25 or 26 or 27 or 28 or 29 or 33 or 34 or 35 or 37 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51 or 53 or 54 or 55 or 56 or 71
74. or/1-22
75. 72 or 74
76. 69 and 73 and 75
77. limit 76 to human

OVID AMED:

1. back pain.mp.
2. backache*.mp.
3. exp Low Back Pain/
4. low back pain.mp.
5. exp Lumbar Vertebrae/
6. exp Spondylolisthesis/
7. ((lumb$ or back) adj pain).ti,ab.
8. exp Sciatica/
9. sciatic*.mp.
10. back-ache.mp.
11. lumbago.mp.

12. exp Intervertebral Disc Displacement/

13. dorsalgia.mp.

14. exp Spondylosis/

15. exp Spondylitis, Ankylosing/

16. exp Spondylitis/

17. spondy*.mp.

18. exp Back Injuries/

19. back injur*.mp.

20. exp “Activities of Daily Living”/

21. activity of daily living.mp.

22. physical function*.mp.

23. limitation of activity*.mp.

24. daily living activity.mp.

25. exp Self Care/

26. physical.mp.

27. performance.mp.

28. abilit*.mp.

29. exp “Recovery of Function”/

30. function.mp.

31. exp Movement/

32. movement.mp.

33. function* task*.mp.

34. exp Mobility Limitation/

35. mobility.mp.

36. function* status.mp.

37. exp Health Status/

38. function* abilit*.mp.

39. ADL.mp.

40. exp Walking/

41. exp Gait/

42. exp Physical Fitness/

43. performance based test*.mp.

44. exp Psychomotor Performance/

45. performance test*.mp.

46. exp Stair Climbing/

47. objective measure*.mp.

48. exp Mobility Limitation/

49. limitation.mp.

50. standing.mp.

51. sitting.mp.

52. exp Community Participation/

53. physical task*.mp.

54. exp Patient Transfer/

55. exp PSYCHOMETRICS/

56. exp “Reproducibility of Results”/

57. valid*.mp.

58. reliab*.mp.

59. reproducib*.mp.

60. repeatability.mp.

61. responsiveness.mp.

62. sensitiv*.mp.

63. specificity.mp.

64. psychometr*.mp.

65. 55 or 56 or 57 or 58 or 59 or 60 or 61 or 62 or 63 or 64

66. exp Spinal stenosis/

67. 27 or 28 or 30 or 32 or 35 or 49

68. 26 and 67

69. 20 or 21 or 22 or 23 or 24 or 25 or 29 or 30 or 31 or 33 or 34 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 50 or 51 or 52 or 53 or 54 or 68

70. or/1-19

71. 66 or 70

72. 65 and 69 and 71

CINAHL:

S69	S57 AND S66 AND S67
S68	S57 AND S66 AND S67
S67	S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23
S66	S58 OR S59 OR S60 OR S61 OR S62 OR S63 OR S64 OR S65

S65	(MH “Sensitivity and Specificity”) OR “sensitiv*”
S64	“responsiveness”
S63	“repeatability”
S62	(MH "Reliability") OR (MH "Reliability and Validity") OR "reliab*"
S61	(MH "Predictive Validity") OR (MH "Discriminant Validity") OR (MH "Criterion-Related Validity") OR (MH "Consensual Validity") OR (MH "Concurrent Validity") OR (MH "Construct Validity") OR "valid*"
S60	(MH "Validation Studies") OR "VALIDATION STUDIES" OR (MH "Predictive Validity") OR (MH "Reliability and Validity") OR (MH "Internal Validity")
S59	(MH “Reproducibility of Results”) OR “Reproducibility”
S58	(MH "Psychometrics") OR "psychometric"
S57	S24 OR S25 OR S26 OR S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42 OR S43 OR S44 OR S45 OR S46 OR S47 OR S48 OR S51 OR S52 OR S53 OR S54 OR S55 OR S56
S56	S27 AND S50
S55	(MH “Leisure Participation (Iowa NOC)”)
S54	(MH “Sports Participation”)
S53	“objective measure”
S52	(MH “Stair Climbing”)
S51	S49 AND S50
S50	“Physical”
S49	(MH "Patient Assessment+")
S48	(MH “Rising”)

S47	(MH "Standing+")
S46	(MH “Sitting”)
S45	(MH “Walking+”)
S44	(MH "Gait+")
S43	(MH "Functional Assessment+")
S42	(MH "Functional Status")
S41	(MH “Ambulation Aids+”)
S40	(MH “Physical Mobility”)
S39	(MH "Structural-Functional-Movement Integration+")
S38	(MH "Movement+")
S37	“Physical Abilit*”
S36	“Physical Conditioning”
S35	(MH “Job Performance”)
S34	(MH “Physical Performance”)
S33	(MH "Motor Activity+")
S32	(MH "Human Activities+")
S31	(MH “Physical Activity”)
S30	(MH “Physical Stimulation+”)
S29	(MH “Functional Status”)
S28	(MH “Physical Mobility”)
S27	"function"
S26	(MH "Self Care+")
S25	“activity of daily life”

S24	(MH “Activities of Daily Living+”)
S23	“back-ache”
S22	“coccy* pain”
S21	“lumbago”
S20	“intervertebral disc degeneration”
S19	(MH “Intervertebral Disk Displacement”)
S18	(MH “Intervertebral Disk+”)
S17	(MH “Sciatica”)
S16	(MH “Sciatic Nerve+”)
S15	(MH “Coccydynia”)
S14	(MH “Spondylolisthesis”)
S13	(MH "Spondylarthritis+")
S12	(MH "Spondylolysis+")
S11	(MH "Spondylosis+")
S10	(MH “Spondylitis, Ankylosing”)
S9	(MH “Spinal Injuries+”)
S8	(MH "Spondylolysis+")
S7	(MH “Osteoarthritis, Spine+”)
S6	(MH “Lumbar Vertebrae”)
S5	“dorsalgia”
S4	(MH “Back”)
S3	“backache”
S2	(MH “Low Back Pain”)

S1	(MH “Back Pain+”)

SPORTDiscus:

S52	(S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10) AND (S50 AND S51)
S51	S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10
S50	S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42 OR S43 OR S44 OR S45 OR S46 OR S47 OR S48 OR S49
S49	(MH “Leisure Participation (Iowa NOC)”)
S48	(MH “Sports Participation”)
S47	“objective measure”
S46	(MH “Stair Climbing”)
S45	(MH "Patient Assessment+")
S44	(MH "Posture+")
S43	(MH “Rising”)
S42	(MH "Standing+")
S41	(MH “Sitting”)
S40	(MH “Walking+”)
S39	(MH "Gait+")
S38	(MH "Functional Assessment+")
S37	(MH "Functional Status")
S36	(MH “Task Performance and Analysis+”)

S35	(MH “Ambulation Aids+”)
S34	(MH “Physical Mobility”)
S33	(MH "Structural-Functional-Movement Integration+")
S32	(MH "Movement+")
S31	“Physical Abilit*”
S30	“Physical Conditioning”
S29	(MH “Exercise Test+”)
S28	(MH “Job Performance”)
S27	(MH “Physical Performance”)
S26	(MH "Motor Activity+")
S25	(MH “Human Activities+”)
S24	(MH “Physical Activity”)
S23	(MH “Physical Stimulation+”)
S22	(MH "Functional Status")
S21	"function"
S20	(MH “Quality of Life+”)
S19	(MH "Self Care+")
S18	“activity of daily life”
S17	(MH “Activities of Daily Living+”)
S16	“back-ache”
S15	"coccy* pain"
S14	“lumbago”
S13	“intervertebral disc degeneration”

S12	(MH “Intervertebral Disk Displacement”)
S11	(MH “Sciatic Nerve+”)
S10	(MH “Spondylitis, Ankylosing”)
S9	(MH “Spinal Injuries+”)
S8	(MH “Osteoarthritis, Spine+”)
S7	(MH “Lumbar Vertebrae”)
S6	“dorsalgia”
S5	(MH “Back”)
S4	(MH “Back Injuries+”)
S3	“backache”
S2	(MH “Low Back Pain”)
S1	(MH “Back Pain+”)

2.6.2 Appendix B: Inclusion and Exclusion Criteria Form

Inclusion/Exclusion Criteria for studies that developed or used Physical Performance-Based measures to assess physical function in patients with low back pain (LBP)	
Reviewer Name:	**Date:**
Author Name: **Study ID:**	**Year:**
Title:	**Journal:**

Study Design:	**Included**	**Comments**
	□ Cohort Study □ Cross Sectional □ Delphi □ Clinical Trails	
Participants:	**Included**	**Comments**
	□ LBP	
Level of measurements:	**Included**	**Comments**
	□ Physical Performance-Based Measures	
Outcome:		
	□ Physical function outcome measures.	

This table will be updated for any additional reasons of exclusion criteria identified during the screening and searching process.

2.6.3 Appendix C-1: Included Studies Data Extraction Sheet

Author	Study Design	Study Language	Study Objective	Inclusion/Exclusion Criteria	Participants Demographic Information (Age, Sex)	Sample Size	Physical Performance-based Tests (PBMs)	Description of PBMs used		Construct	Psychometric Properties Evaluated
								Content	Scoring		

2.6.3 Appendix C-2: Summary of the Psychometric properties of outcome measures

Outcome Measure	Back Pain Classification	Psychometric Property Evaluated										
		Reliability				Validity						Responsiveness
		Test-Retest	Inter-Rater	Intra-Rater	Internal Consistency	Criterion Validity		Face & Content Validity	Construct Validity			Criterion & Construct approaches
						Predictive	Concurrent		Convergent	Known-Groups	Discriminant	

2.6.4 Appendix D-1: Result Rating cut-off points per psychometric properties.

Psychometric Properties	Rating	Result Rating Criteria
Reliability	+	ICC or weighted κ ≥ 0.70
	?	ICC or weighted κ not reported
	–	Criteria for "+" not met
Measurement Error	+	SDC or LoA < MIC
	?	MIC not defined
	–	Criteria for "+" not met
Hypothesis testing for construct validity	+	Same constructs: correlation is expected to be ≥ 0.7 Related constructs: correlation is expected to be ≥ 0.5 Unrelated constructs: correlation is expected to be ≥ 0.3
	?	Solely correlations determined with unrelated constructs No correlations with instrument(s) measuring related construct(s) reported No differences between relevant groups reported
	–	Criteria for "+" not met
Criterion Validity	+	Same constructs: correlation is expected to be ≥ 0.7 Related constructs: correlation is expected to be ≥ 0.5 Unrelated constructs: correlation is expected to be ≥ 0.3
	?	Solely correlations determined with unrelated constructs No correlations with instrument(s) measuring related construct(s) reported No differences between relevant groups reported
	–	Criteria for "+" not met
Responsiveness	+	AUC ≥ 0.70 Same constructs: correlation is expected to be ≥ 0.7 Related constructs: correlation is expected to be ≥ 0.5 Unrelated constructs: correlation is expected to be ≥ 0.3
	?	Solely correlations determined with unrelated constructs No correlations with instrument(s) measuring related construct(s) reported No differences between relevant groups reported
	–	Criteria for "+" not met
ICC, Interclass correlation coefficient; κ, Kappa; SDC, Smallest Detectable Change; LoA, Limit of Agreement; MIC, Minimal Important Change; AUC, Area Under the Curve.		

2.6.4 Appendix D-2: GRADE criteria for data analysis of level of evidence per performance-based measurements.

Overall Result Rating per PBM: Positive vs. Negative
Positive Results (+) : 75% of the results in accordance with the Hypotheses in Appendix D-1
Negative Results (–) : Criteria for "+" not met
Criteria for Level of Evidence per PBM:
Strong. Positive result ratings in at least 1 Very Good-quality article or 2 Adequate-quality articles
Moderate. Positive result ratings in at least 1 Adequate-quality article or 2 Doubtful-quality articles
Limited. Negative Rating with at least 1 Very Good-quality article Negative Rating with at least 1 Adequate-quality article Positive or Negative Ratings of 1 Doubtful-quality article
Poor. All eligible articles were of Inadequate-quality articles

Chapter Three: A Systematic Review of the psychometric Properties of Performance-Based Measures to assess Physical Function in Low Back Pain patients.

Title: A Systematic Review of the psychometric Properties of Performance-Based Measures to assess Physical Function in Low Back Pain patients.

Authors:
Maysa Alnattah, PT BSc, MS MSc, PH PgD, School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Luciana G. Macedo, PT Ph.D., School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Ayse Kuspinar, PT Ph.D., School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Marla Beauchamp, PT Ph.D., School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada

Corresponding author:
Maysa Alnattah, alnattam@mcmaster.ca, room 308, IAHS, McMaster University, Hamilton, ON, Canada

3.0 Abstract

Background: Physical function is an important core outcome in low back pain (LBP) that is primarily assess by Patient-reported outcome measures (PROMs). However, physical performance should also be assessed using performance-based measures (PBMs). Previous systematic reviews documented the psychometric properties of PBMs in LBP but were not comprehensive. The purpose of this study was to identify PBMs developed for or used to assess physical function in LBP population and to systematically review studies evaluating the psychometric properties of these PBMs.

Methods: Five databases were searched (MEDLINE, EMBASE, AMED, CINAHL, and SPORTDiscus) using search terms involving domains of LBP, performance tests/measurers, physical function, and psychometric properties. Studies were included if they recruited individuals with LBP (with no serious pathologies), used PBMs to assess physical function, and investigated any psychometric properties of these Measurements. Two authors completed study screening, evaluation, and data extraction. Data synthesis was based on a pre-established criterion for results rating and the COSMIN (COnsensus-based Standards for the selection of health status Measurement INstruments) Risk of Bias checklist 2018 (COSMIN-ROB).

Results: There were 47 studies that met the inclusion criteria with five LBP diagnosis (e.g., non-specific LBP) and different LBP durations (e.g., acute, chronic). In general, findings included 115 PBMs. Most of the level of evidence were generated from single studies for each PBM or psychometric property. The majority of the included studies had high risk of bias assessed by COSMIN-ROB checklist. Large number of studies did not find PBMs to have good psychometric properties as results/scores did not meet the pre-defined thresholds/hypothesis for good psychometrics. The great majority of PBMs' psychometric properties were found to have low level of evidence.

Conclusion: There is a significant heterogeneity of studies evaluating the psychometric properties of PBMs used to assess physical function in LBP patients leading to limited level of evidence. Therefore, such PBMs need to be used with great cautious. Moreover, there is a large need for more high-quality studies that investigate psychometric properties PBMs of physical function in LBP.

3.1 Introduction

Low Back Pain (LBP) is the leading cause of disability world-wide.[1] LBP related disability is linked to increased demands on the health care system and the economy.[1] In Canada, the total annual LBP-related medical cost estimates (Direct-Costs) ranges from $6 to $12 billion, not including societal costs associated with disability payment and work loss productivity (Indirect-Costs).[2]

Clinical assessment is an important and critical step in both research and evidence-based clinical practice.[3] A recently published Delphi study identified and recommended three Core Domains Set (COS) of outcomes to be assessed in LBP trials: physical function, pain intensity, and health-related quality of life (HRQoL).[4] All outcome measures that were recommended within each domain, especially physical function, were self-reported measures.[5] This is because the previous Delphi study only selected PROMs given their feasibility and because they are the most frequently used and recommended measurements in the LBP literature.[4] Patient-Reported Outcome Measures (PROMs) have many advantages such as being easy to administer, inexpensive, and able to provide patients' perceptions of disability.[5] Nevertheless, PROMs suffer from significant limitations such as recall bias, social desirability, errors in self-observation, and misinterpretation of terminology.[6-12] In addition to these limitations, recent systematic review have identified that currently available PROMs have considerable psychometric limitations in terms of content and structural validity, thus adding to the challenge of using these outcomes within research and clinical practice.[13 14]

Given the limitations associated with PROMs, it is suggested that Performance-Based Measures (PBMs) should be used in addition to PROMs when assessing physical function.[5 15-18] However, PBMs are not currently listed within the core outcome set recommended outcome measures for LBP.[14] This is partially due to the lack of evidence for the selection of the most appropriate tests as well as their psychometric properties. To our knowledge, two systematic reviews on the psychometric properties of PBMs used to assess physical function in LBP patients have been published.[19 20] A review published in 2018 investigated the reliability of 38 PBMs for LBP in 20 studies (search dated on June 24, 2017),[19] and a review published in 2019 provided a more comprehensive investigation of the reliability, validity and responsiveness of 18 PBMs and 25 studies (search dated on August 29, 2018).[20] Although these reviews provided promising results on the reliability and validity of some PBMs, and the reviews were generally well conducted, they have several limitations including: 1) inconsistent criteria for the definition of physical function; 2) exclusion of battery tests and Functional Capacity Evaluation Forms, 3) the use of an outdated version of the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist for methodological quality, 4) the exclusion of acute LBP patients, and 5) only including English language studies.[19 20]

Thus, the purpose of the current systematic review was to address the shortcomings of the previously published reviews by performing a comprehensive search to (1) identify all PBMs that have been developed or used to assess physical function in LBP patients; and (2) synthesize the available evidence on the psychometric properties (validity, reliability and responsiveness) of those identified physical PBMs using the most up to date systematic review standards.

3.2 Method

Study design:

A systematic review of PBMs that were developed or used to assess physical function in patients with LBP with a focus on psychometric properties (reliability, validity and responsiveness). The review followed the updated COSMIN systematic review methodology manual 2018 for conducting the review, assessing the included studies' methodological quality, and data synthesis of level of evidence.[21] Also, some of the recommendations of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for systematic review reporting was followed when necessary (e.g., piloting data extraction forms).[22]

Search:

A search was conducted to identify: 1) studies that reported on the development or use of physical function PBMs for LBP patients, and 2) studies that evaluated the psychometric properties of physical function PBMs in patients with LBP. A flow diagram illustrated the search process in Figure 1. Studies that have evaluated the psychometric properties of outcome measures in LBP population, even if the tool was not developed for LBP, were included.

The following databases were used: MEDLINE (OvidSP, 1946 to 19 May 2019); EMBASE (OvidSP, 1980 to 19 May 2019); AMED (OvidSP, 1985 to 19 May 2019); CINAHL (EBSCO, 1981 to 19 May 2019); and SPORTDiscus (EBSCO, 1800 to 19 May 2019). The search terms that were used spanned the following domains: low back pain, physical performance-based measures, physical function, and psychometric properties. Once we identified outcome measures from the previous search, we also did a second search using the name of the identified PBMs and psychometric property terms. See appendix (A) for search terms. Hand searches of reference lists of similar systematic reviews and all included studies was conducted. We also conducted citation tracking of included studies using ISI Web of Science.

Eligibility Criteria:

- **Type of studies**:

Studies that developed or used at least one PBM of physical function in LBP were considered. We included studies that tested the psychometric properties of the physical PBMs identified. No restriction of language and publication date were imposed. We did not impose limits to study designs either.

- **Type of participants**:

Studies that have included participants with age of 18 years old and above, and any sex/gender who have LBP were considered. In this review, LBP is defined as pain or discomfort in the lower back region that is attributed to any known or unknown pathoanatomical cause.[23-26] Studies that have included participants with specific and non-specific low back pain of any duration were included. Studies on the development of a PBM of physical function that used mixed populations in relation to diagnosis (e.g., OA) were only included if data about low back pain patients was available. Authors of included studies that used populations with mixed specific and non-specific low back pain patients were contacted to ask for data of different sub-groups, however, if data was not available it was excluded. Studies that included LBP due to the following serious pathologies were excluded: spine deformity (Scoliosis), cancer, fractures, inflammations, etc. Pregnancy was also excluded.

- **Types of outcome**:

Physical Performance-Based Measures (PBMs) that were intended to measure physical function in LBP patients were considered. PBMs are measures and tools that health professionals and researchers use to collect information about patients' current physical presentations. They are used to assess, measure and/or observe patients' actual performance according to instructions on a set of functional tasks, including but not limited to: 1) self-care skills, 2) transfers, 3) mobility, and 4) movements.[27 28]

Physical function was defined according to the International Classification Of Functioning Disability and Health (ICF), which mostly falls under the umbrella term of Activity.[29] PBMs of physical function were considered as tests of Activity if they were used to determine patients' ability to execute a specific physical task or action in a safe and timely manner.[29] Examples of such measures are the Timed–Up and Go Test, Dynamic Gait Index, and Tinetti Mobility Test. PROMs and impairment-based tests (e.g., Range of Motion, muscle performance, etc.) were excluded.

We also reported the psychometric properties of identified PBMs. The three domains of the psychometric properties that were included were reliability, validity and responsiveness as defined by COSMIN.[21]

Screening:

Screening of titles and abstracts, and eligible studies' full texts, was conducted independently by two reviewers. An inclusion form developed prior to the start of the review was used to assess the inclusion eligibility. A third reviewer resolved any disagreements.

Risk of bias:

The COSMIN Risk Of Bias checklist of 2018 (COSMIN-ROB) was used to assess the risk of bias of the included studies.[30] Two reviewers independently evaluated the included studies. A third reviewer resolved any disagreements. Prior to data collection, authors extracting data participated in an informal training on the use of the COSMIN-ROB checklist as recommended by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[31]

Data extraction and analysis:

Two reviewers extracted data independently into a data extraction form developed and pilot tested before the beginning of the study. All tests of reliability, validity, and responsiveness that were used to evaluate the PBMs were considered. Reliability tests included test-retest, inter-rater, and intra-rater. Validity tests included construct validity (convergent and known-group). Responsiveness included the criterion and construct approaches.

Data synthesis was based on two levels of assessment: 1) COSMIN-ROB checklist, and 2) rating of psychometric results. Result rating was carried out by comparing the psychometric property score to a pre-determined cut-off point or hypothesis (see Table 1). Using both the COSMIN-ROB checklist and results rating assessment, data synthesis was conducted using the GRADE criteria (Grading of Recommendations, Assessment, Development and Evaluations) in Table 2.[21] For example, a PBM that had a grade Very Good on COSMIN-ROB and exceeded the predetermined cut-off point for test-retest reliability (ICC ≥ 0.70) received a level of evidence of "Strong". However, if the test had an Inadequate grade or did not exceed the cut-off point, the level of evidence was considered to be "Poor". The order of grades from highest to lowest was as follows: Strong, Moderate, Limited, and Poor.

Table 1: Result rating criteria per psychometric properties

Psychometric Properties	Rating	Result Rating Criteria
Reliability	+	ICC or weighted $\kappa \geq 0.70$
	?	ICC or weighted κ not reported
	–	Criteria for "+" not met
Measurement Error	+	SDC or LoA < MIC
	?	MIC not defined
	–	Criteria for "+" not met
Hypothesis testing for construct validity	+	Same constructs: correlation is expected to be ≥ 0.7 Related constructs: correlation is expected to be ≥ 0.5 Unrelated constructs: correlation is expected to be ≥ 0.3
	?	Solely correlations determined with unrelated constructs No correlations with instrument(s) measuring related construct(s) reported No differences between relevant groups reported
	–	Criteria for "+" not met
Criterion Validity	+	Same constructs: correlation is expected to be ≥ 0.7 Related constructs: correlation is expected to be ≥ 0.5 Unrelated constructs: correlation is expected to be ≥ 0.3

	?	Solely correlations determined with unrelated constructs No correlations with instrument(s) measuring related construct(s) reported No differences between relevant groups reported
	–	Criteria for "+" not met
Responsiveness	+	AUC ≥ 0.70 Same constructs: correlation is expected to be ≥ 0.7 Related constructs: correlation is expected to be ≥ 0.5 Unrelated constructs: correlation is expected to be ≥ 0.3
	?	Solely correlations determined with unrelated constructs No correlations with instrument(s) measuring related construct(s) reported No differences between relevant groups reported
	–	Criteria for "+" not met
ICC, Interclass correlation coefficient; κ, Kappa; SDC, Smallest Detectable Change; LoA, Limit of Agreement; MIC, Minimal Important Change; AUC, Area Under the Curve.		

Table 2: GRADE synthesis criteria of four levels of evidence used to evaluate PBMs.

Overall Result Rating per PBM:
Positive Results (+) : 75% of the results in accordance with the Hypotheses in table 1.
Negative Results (–) : Criteria for "+" not met
Criteria for Level of Evidence based on number of studies and COSMIN-ROB
Strong. Positive result ratings in at least 1 Very Good-quality article or 2 Adequate-quality articles
Moderate. Positive result ratings in at least 1 Adequate-quality article or 2 Doubtful-quality articles
Limited. Negative Rating with at least 1 Very Good-quality article Negative Rating with at least 1 Adequate-quality article Positive or Negative Rating of 1 Doubtful-quality article
Poor. All eligible articles were of Inadequate-quality articles

3.3 Results

Studies' search and selection:

The main search and hand search produced 11,053 and 5,959 studies after removal of duplicates, respectively. After title and abstract screening, a total of 292 studies were identified for full text screening. In addition, there were 63 articles identified from a subsequent Web of

Science search that was included with the full-text screening. In total, 47 studies were included in this systematic review.[32-78] The reasons for exclusion of full texts are summarized in Figure 1.

Studies' characteristics:

Of the 47 included studies,[32-78] there were different types of LBP included: non-specific LBP (17 studies),[32 33 38 43 44 51 52 60 63 64 66-68 71 72 76 78] mixed population in terms of diagnosis (14 studies),[34 35 37 40 41 45 55-59 61 62 77] Lumbar Spinal Stenosis (12 studies),[36 39 42 46-49 53 65 73-75] LBP due to degenerative changes (3 studies),[50 69 70] and muscular related LBP (1 study).[54] There were also 14 studies that did not provide a cause LBP for the included patients.[34 35 37 40 41 45 55-59 61 62 77] Most of the studies included LBP of chronic duration (43 studies),[32-44 46-61 63-69 71-75 77 78] one study included participants with acute LBP,[76] and three included a population of mixed LBP duration (acute, subacute, and chronic).[45 62 70] See Table 3 for a summary of the included studies' characteristics.

Risk of bias was assessed using the COSMIN-ROB checklist. Six studies had a Very Good quality rating,[37 47 50 66 73 78] two had an Adequate rating,[38 69] 18 had a Doubtful rating,[32 35 39 41 43 45 51 52 54-56 58 62 64 68 72 74 75] and the remaining 21 studies had an Inadequate quality rating.[33]

In total, there were 115 PBMs included in the review. The included PBMs involved tests that are categorized as follow: Walking tests (37 PBMs),[32-35 37 39-41 43 48 50 53 60 61 65-68 71-74 76 79] Battery Tests (16 PBMs),[33 38 44 45 54 56 59 64 66 68 69 71 76 78] Lifting tests (11 PBMs),[38 47 55 56] Treadmill walking tests (7 PBMs),[36 42 46 49 65 75 77] Sit-to-Stand tests (11 PBMs),[32 33 40 41 52 60 61 67 76] Stair Stepping tests (7 PBMs),[33 50 55] Balance tests (5 PBMs),[51 53 57 58 62] and other functional tests (21 PBMs).[38 40 55 56]

Reliability

Test-Retest Reliability

A total of 13 studies investigated the test-retest reliability of 15 PBMs.[32 35 41 42 45 54 56 64 68 70-72 75] The highest level of evidence for test-retest reliability was "Moderate" for two PBMs: 50-Feet Walk Test (s) and Back Performance Scale.[32 33 41 45 56 71] This "Moderate" level of evidence was generated from two or three studies for the two PBMs.[32 33 41 45 56 71] "Limited" level of evidence was found for 12 PBMs among which eight PBMs had a positive result rating (ICC or Kappa ≥ 0.70) and five PBMs had a negative result rating (ICC or Kappa ≤ 0.70).[35 45 54 68 75] The remaining PBMs received "Poor" level of evidence, due to the very low score on COSMIN-ROB, for positive rating (ICC or Kappa ≥ 0.70) and were generated from a single study per test.[42 70 71] See Table 4 for more details about PBMs' test-retest reliability.

Intra-Rater Reliability

Eight studies investigated the intra-rater reliability of 18 PBMs.[32 51 52 55 57 62 69 80] All PBMs were investigated by a single study except for one PBM (One Leg Stand Test) which was investigated by three studies.[57 62 80] Only two PBM (Functional Capacity Evaluation-Safe Maximum Lifting and Single Leg Stance Test) received a "Moderate" level of evidence.[57 62 69 80] The other 16 PBMs had limited levels of evidence among which five PBMs had negative result ratings (ICC or Kappa ≤ 0.70),[55] and 11 received positive result ratings (ICC or ≥ 0.70).[32 51 52] See Table 5 for more details about PBMs' intra-rater reliability.

Inter-Rater Reliability

A total of seven studies investigated the inter-rater reliability of 21 PBMs.[32 47 55 57 69 78 81] Six PBMs (Isernhagen Work Systems Functional Capacity Evaluation-Lifting Test (Borg CR-10 scale), Back-Torso Lift Test, Shoulder Lift Test, Carrying Lifting Strength Test, Lower Lifting Strength Test, and Upper Lifting Strength Test) had a "Strong" level of evidence; however, these were generated by a single study for each PBM.[47 78] Sixteen PBMs received a "Limited" level of evidence which was also generated by a single study for each PBM,[32 47 55 62 69 78] except for one PBM (One Leg Stance) that received a "Limited" level of evidence generated from two studies (one study had an Inadequate score with positive result rating,[57] and one had a Doubtful score with negative result rating.)[62] Among all the PBMs that received a "Limited" level of evidence, four PBMs had positive results ratings (ICC or ≥ 0.70),[32 62] and 12 PBMs received a negative result rating (ICC or Kappa ≤ 0.70).[55 57 69 78] See Table 6 for more details about PBMs' inter-rater reliability.

Measurement Error

Data on measurement error was poorly reported. Only two studies out of 47 had data on smallest detectable change (or limit of agreement) and minimal detectable change.[33 71] Both studies had negative result rating (did not meet the hypotheses for good measurement error).[33 71]

Hypothesis Testing–Construct Validity

Convergent Validity

Convergent validity was the most commonly assessed psychometric property (24 studies).[34 36-41 45 46 48 49 52 53 56 59-61 65-67 70 71 73 77] Out of these 24 studies, 30 PBMs were investigated and correlated to 35 self-reported questionnaires or questions (e.g., ODI, Self-Estimated Walking Distance). Nine PBMs received a "Strong" level of evidence that was generated from a single high-quality study for each (Shuttle Walk Test, Ambulatory-Treadmill Test (Distance), Treadmill Tolerance Test (Time), Timed Up-and-Go Test, 5-Repetition Sit To Stand (Time), Back Performance Scale, Functional Capacity Evaluation, and Functional Test Index);[36-38 49 53 66] and three studies for one PBM (Self-Paced Walking Test (Distance)).[39 65 73] A "Moderate" level of evidence was found for three PBMs (Motorised Treadmill Test (Distance and Time), Free Walking Velocity Test, and Self-Paced Walking Test (Distance or Time)) which was generated from a single study for each PBM.[48 65] Three PBMs received "Limited" level of evidence, because of the negative result ratings ($r \leq 0.50$), generated from 2-5 studies.[37 40 41 56 60 61 66 67 71] The remaining PBMs received "Limited" level of evidence generated from a single study for each PBM with negative result ratings ($r \leq 0.50$).[38 40 46 48 52 53 59 61 66 71 77] See Table 7 for more details about PBMs' convergent validity.

Known-Groups Validity

Only two studies investigated the known-groups validity of seven PBMs. [45 63] One study of high quality (Very Good score on COSMIN-ROB) investigated the ability of six PBMs to discriminate between patients with LBP and patients without LBP. [62] All six PBMs (20-Steps Stair Climbing Test (Time), Roll-Up Test, Stand-to-Floor Test, Sock Test, Pick-Up Test, and 5-Repetition Lift Test (Ordinal 0-3)) had a positive result rating (difference between two groups is significant (p-value ≤ 0.05). [62] Therefore, they received a "Strong" level of evidence.

The second study had a Doubtful score on the COSMIN-ROB. [44] It investigated the Back Performance Scale's ability to discriminate between 3 subgroups reporting their pain level (high pain NPS ≥ 4 vs. low pain NPS < 4), activity level (high activity: not reduced or slightly reduced vs. low activity: fairly or very reduced) and work status (employed vs. on sick leave). [44] The Back Performance Scale was able to discriminate between patients who reported different scores on pain level and activity level (positive result rating) but not work status (negative result rating). [44] The overall level of evidence that was generated for this PBM was "Limited" due to the low quality of the study. [44] For more details about the PBMs' known-group validity, see table 8.

Responsiveness

Responsiveness - Construct Approach (hypothesis testing: comparison with other PROMs)

Seven studies investigated the responsiveness of 18 PBMs using a construct approach in which they investigated the correlation between these measures with PROMs.[34 38 44 50 65 71 74] Only one PBM (Self-Paced Walking Test (Distance)) received a “Strong” level of evidence; however, this was generated from one single study that only included patient with lumbar spine stenosis.[74] The remaining PBMs (n=17) received a “Limited” level of evidence investigated by a single study per each PBM, with three PBMs receiving a positive result rating ($r \geq 0.50$),[38 65] and 15 PBMs receiving a negative result rating ($r \leq 0.50$).[34 38 44 50 65 71] For more details about the PBMs’ responsiveness–hypothesis testing and comparison with other PROMs, see table 9.

Responsiveness - Construct Approach (hypothesis testing: comparison between subgroups)

Only two studies investigated the responsiveness (comparison between subgroups) of five PBMs. [44 71] One study of high quality (Very Good score on COSMIN-ROB) investigated the ability of four PBMs to discriminate between patients who rated themselves as very much improved, much improved, slightly improved or had no change. [70] Two PBMs (Lift Test and Back Performance Scale) had a positive result rating (able to discriminate) and two PBMs (Progressive Isoinertial Lifting Evaluation and 15-m walk test (m/s)) had a negative result rating (not able to discriminate). [70]

The other study had a Doubtful score on the COSMIN-ROB. [43] It investigated the Physical Work Performance Evaluation (Overall score) ability to discriminate between 2 subgroups (better score vs. worse score) reporting their improvement using a 15-point scale ranging from -7 to +7. [43] The test had a negative result rating, and the overall level of evidence is “Limited”. [43] See table 10 for more details of the PBMs’ responsiveness (comparison between subgroups) using a construct approach.

Responsiveness - Criterion Approach (comparison to a gold standard)

There were six studies that investigated responsiveness of 11 PBMs using a criterion approach.[33 37 50 65 71 74] All of these studies used a Generic-Global Rating Scale (generic-GRS) to discriminate between patients who improved or were unchanged.[33 37 50 65 71 74] In addition, two studies used a specific-GRE of physical function to evaluate five PBMs.[50 74] Specific-GRE include patient-reported question that were specific to physical function as opposed to generic GRS that were general to the condition and not physical function.

Generic-Global Rating Scale (Unchanged-Improved):

Five studies that used a generic-GRS had similar GRS-questions; meaning, patients were asked to score whether their disability/function levels were improved or unchanged after treatment.[33 37 50 65 71] Areas under the curve (AUC) ranged from 0.545 to 0.77 for all PBMs. A "Strong" level of evidence was found for five PBMs (1-Minute Stair Climbing Test, 5-Minute Walk Test, 5-ft Walk Test, Time Up-and-Go Test, Shuttle Walk Test (Distance)); however, this was generated from a single study for each PBM.[37 50] Four PBMs received a "Limited" level of evidence that was also generated from a single study for each PBM.[65 71] Among these five PBMs, two had a positive result rating (AUC ≥ 0.70) and three had negative result rating (AUC ≤ 0.70).[65 71] The remaining PBMs received a "Poor" level of evidence due to their very low score (Inadequate) on the COSMIN-ROB.[33] See table 11 for more details of the PBMs' responsiveness using Generic-GRS.

Specific-Global Rating Scale (Unchanged-Improved):

Two studies used specific-GRS scales to investigate the criterion responsiveness of five PBMs.[50] A "Strong" level of evidence was found for two PBMs (1-Minute Stair Climbing Test, Time Up-and-Go Test,).[50] A "Limited" level of evidence was found for two PBMs (one had negative result ratings (AUC < 0.70,[50] and the other had positive result rating (AUC ≥ 0.70).[74] See table 12 for more details of the PBMs' responsiveness using Specific-GRS.

Table 13 summarizes all the evidence of the evaluated psychometric properties for each PBM per LBP population. The data presented were for the PBMs that were evaluated by two or more studies, unless only a single study existed.

3.4 Discussion and conclusion

There were 47 studies included in this systematic review evaluating a total of 115 PBMs.[32-78] Overall, studies were widely heterogeneous, evaluated different populations in terms of diagnosis, different types of PBMs, and different psychometric properties. The majority of the evidence for each PBM was limited to 1 or 2 psychometric properties, and evidence was mostly derived from single studies with overall low quality. Almost half of the studies' results did not meet this review's pre-determined hypotheses for good psychometric testing. There was not a single PBM that was tested for all psychometric properties (reliability, validity and responsiveness) or that when tested was found to have strong level of evidence. In general, most of the included PBMs had limited level of evidence for their psychometric properties.

Among the evaluated PBMs for non-specific LBP, the 50-feet Walk Test and the 5-repetition Sit-to-Stand Test (time) were the ones that were most comprehensively evaluated. However, the highest level of evidence identified for these two tests was "Moderate", and this was only for test-retest reliability (ICC ≥ 0.70) of 50-feet Walk Test. For validity and responsiveness, both tests showed "Limited" and "Poor" level of evidence, respectively. Therefore, the outcomes of these tests should be interpreted with great caution as it is unclear how much measurement error is included in these measures as well as whether measures trully reflect physical function. For Lumbar Spinal Stenosis, only the Self-Paced Walk Test (Distance) showed high level of evidence for convergent validity and responsiveness; however, there is no data on its reliability. For all other diagnosis, the 50-ft Walk Test (s), Timed Up and Go Test (s) and Back Performance Scale had "Strong" level of evidence on many but not all of the psychometric properties. In general, the above mentioned PBMs are promising tests that can be combined to PROMs when assessing physical function in LBP trials but need more investigation.

The results of this review often demonstrated more promising validity and responsiveness results for lumbar spinal stenosis as opposed to nonspecific low back pain. Patients with lumbar spinal stenosis often reporting neurogenic claudication with difficulty with walking as a primary complaint. Therefore, it is to expect that a walking test would reflect these patient's physical function. However, there is greater heterogeneity in the presentation of patients with non-specific LBP which may reflect on poorer psychometric properties overall including ceiling effects.

In general, most of the included studies met the review hypothesis for good reliability; however, they had Doubtful or Inadequate scores on COSMIN-ROB leading to low levels of evidence. High risk of bias was primarily associated with long interim periods between the first and second assessment, and poor description of test conditions (e.g., type of administration, environment, instructions). In a few of the included studies, reliability was not the primary goal and was evaluated using either data from participants that self-identified as not having changed or a small subpopulation of the study. Future high-quality studies have the potential to provide stronger levels of evidence for reliability given that the results of this review were promising, with often moderate to high ICC values.

Convergent validity was the most investigated type of validity. Validity studies demonstrated better quality on the COSMIN-ROB compared to reliability studies. Nevertheless, more than half of PBMs did not meet the review hypothesis for adequate validity. PROMs were commonly used to measure convergent validity with PBMs. Interestingly, the correlations between two similar PBMs (e.g., two different walking tests) often did not meet the hypothesis for good convergent validity. This could be due to having a higher threshold for a positive result rating ($r \geq 0.70$) when assessing constructs that were similar. Regardless, these results also raise questions about the potential various factors that may affect PBMs' performance such as the impact of psychosocial factors on the results (e.g., fear of movement) or even the different components involved in the completion of a PBM (e.g., balance and mobility).

In this review there were a limited number of studies that evaluated responsiveness of PBMs. Only 12 studies investigated responsiveness (five studies used the criterion approach, and seven used the construct approach). In general, the great majority of responsiveness studies had poor quality (low score on COSMIN-ROB). A reason that often contributed to this low score on COSMIN-ROB was the inappropriate time interval between the first and second assessment, where participants' status changed due to factors other than receiving a treatment. In addition, many PBMs' did not meet the threshold of good responsiveness (construct approach: $r < 0.50$; or criterion approach: AUC < 0.70). These results are concerning and demonstrate the need for higher quality evidence before PBMs can be used to evaluate change in physical function in intervention studies or clinical practice.

As aforementioned, there are two previously published systematic reviews on the psychometric properties of PBMs for physical function in LBP. In this review, we identified an additional 97 PBMs, likely due to a more extensive and updated literature search and potentially broader inclusion criteria (e.g., inclusion of treadmill tests and functional capacity tests). There were two other main differences between the previously published reviews and this review: the use of an updated version of the COSMIN-ROB checklist and how hypotheses were generated for results rating. When using the updated COSMIN-ROB 2018 checklist, study scores are not affected by whether the original study had an a-priori hypothesis or low sample size. Thus, in this review, studies often had lower ROB scores leading to higher level of evidence. In contrast, in the review by Jakobsson et al. hypotheses were determined through "derivation" leading to different, often 'easier' thresholds for results ratings. This means that even when the same studies are included, conclusions can be different between the reviews.

The results of the current review are similar to the previous two reviews in terms of reliability, although we identified seven additional studies on inter- and intra-rater reliability. The largest difference between this review and the one by Jakobsson et al. was with respect to validity, with an extra 11 studies included in our review. In the Jakobsson et al. review, most of the results receive d positive ratings; however, in our review, half of the results received negative ratings. This means that our hypotheses' threshold for validity, developed from the updated COSMIN systematic review methodology manual were harder to meet. Further, there were no differences between the results of this review and the previous review on responsiveness. We were able to identify only one new study reporting on this psychometric property.

Strength and limitations

A limitation of this review was the exclusion of studies (3 studies) in which the description of the protocols of the PBMs protocols was very poor, and no measurement unit or total score was presented (e.g., kg, meter, seconds), which may have led to the exclusion of potentially relevant measures. Further, although we included non-English studies (3 studies), our translation was primarily based on informal methods such as www.translate.google.com which we acknowledge may have introduced potential errors in translation.

Important strengths of this review include the use of the updated COSMIN systematic review methodology manual (2018) which recommends the inclusion of eligible studies even if they present with high risk of bias. The previous version recommended the removal of such studies from data synthesis.[21] Further, the updated manual recommends the development of pre-defined hypotheses for each psychometric property (mainly for hypothesis testing and responsiveness) which allows for standardization of ratings and minimizes bias that may have been present in the original study. Other strengths of this study were the use of the global WHO-ICF model to define physical function with consideration of the overlap g between the *Activity* and *Participation* domains; the identification of different types of LBP diagnosis;[21] as well as the inclusion of non-English studies and grey literature.

Directions for future practice

There is limited evidence on the psychometric properties of PBMs used to assess physical function in LBP.[19 20] Therefore, caution is recommended when interpreting PBMs' outcomes to assess physical function in LBP. More high-quality studies evaluating all psychometric properties of PBMs in patients with different LBP diagnoses are needed before these measures can be widely implemented in clinical practice and research. In general, Self-Paced Walk Test (Distance) for lumbar spinal stenosis and the 50-ft Walk Test (s) and Back Performance Scale for other LBP diagnosis are promising PBMs that need more investigation across different LBP diagnoses.

3.5 Reference

1. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* 2018;392(10159):1789-858.
2. Canada BaJ. Low Back Pain http://boneandjointcanada.com2014 [updated 2014. Available from: http://boneandjointcanada.com/low-back-pain/.
3. Wittink H, Nicholas M, Kralik D, et al. Are we measuring what we need to measure? *The Clinical journal of pain* 2008;24(4):316-24.
4. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain* 2018;159(3):481-95.
5. Thonnard JL, Penta M. Functional assessment in physiotherapy. A literature review. *Europa Medicophysica* 2007;43(4):525-41.
6. Stone AA, Shiffman S, Schwartz JE, et al. Patient compliance with paper and electronic diaries. *Controlled clinical trials* 2003;24(2):182-99.
7. Buer N, Linton SJ. Fear-avoidance beliefs and catastrophizing: occurrence and risk factor in back pain and ADL in the general population. *Pain* 2002;99(3):485-91.
8. Heneweer H, Picavet HSJ, Staes F, et al. Physical fitness, rather than self-reported physical activities, is more strongly associated with low back pain: evidence from a working population. *European Spine Journal* 2012;21(7):1265-72.
9. Anderson B, Lygren H, Magnussen LH, et al. What Functional Aspects Explain Patients' Impression of Change after Rehabilitation for Long-lasting Low Back Pain? *Physiotherapy Research International* 2013;18(3):167-77.
10. Guildford BJ, Jacobs CM, Daly-Eichenhardt A, et al. Assessing physical functioning on pain management programmes: the unique contribution of directly assessed physical performance measures and their relationship to self-reports. *British journal of pain* 2017;11(1):46-57.
11. Tucker JM, Welk GJ, Beyler NK. Physical activity in US adults: compliance with the physical activity guidelines for Americans. *American journal of preventive medicine* 2011;40(4):454-61.

12. Mcloughlin MJ, Colbert LH, Stegner AJ, et al. Are women with fibromyalgia less physically active than healthy women? *Medicine and science in sports and exercise* 2011;43(5):905.
13. Chiarotto A. Patient-Reported Outcome Measures: Best Is the Enemy of Good (But What if Good Is Not Good Enough?). *The Journal of orthopaedic and sports physical therapy* 2019;49(2):39-42.
14. Chiarotto A, Ostelo RW, Boers M, et al. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. *Journal of Clinical Epidemiology* 2018;95:73-93.
15. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine* 2011;36(21 SUPPL.):S54-S68.
16. Taylor AM, Phillips K, Patel KV, et al. Assessment of physical function and participation in chronic pain clinical trials: IMMPACT/OMERACT recommendations. *Pain* 2016;157(9):1836-50.
17. Carey TS, Mielenz TJ. Measuring outcomes in back care. *Spine* 2007;32(11 SUPPL.):S9-S14.
18. Turk DC, Melzack R. Trends and future directions in human pain assessment. 2001
19. Denteneer L, Van Daele U, Truijen S, et al. Reliability of physical functioning tests in patients with low back pain: a systematic review. *Spine Journal* 2018;18(1):190-207.
20. Jakobsson M, Gutke A, Mokkink LB, et al. Level of Evidence for Reliability, Validity, and Responsiveness of Physical Capacity Tasks Designed to Assess Functioning in Patients With Low Back Pain: A Systematic Review Using the COSMIN Standards. *Physical Therapy* 2019;99(4):457-77.
21. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018;78:1.
22. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS medicine* 2009;6(7):e1000100.
23. Savigny P, Watson P, Underwood M, et al. Early management of persistent non-specific low back pain: summary of NICE guidance. *BMJ* 2009;338:b1805.

24. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet* 2017;389(10070):736-47.

25. Hoy D, Brooks P, Blyth F, et al. The epidemiology of low back pain. *Best practice & research Clinical rheumatology* 2010;24(6):769-81.

26. Freburger JK, Holmes GM, Agans RP, et al. The rising prevalence of chronic low back pain. *Archives of internal medicine* 2009;169(3):251-58.

27. de Vet HCW, Terwee CB, Mokkink LB, et al. Measurement in Medicine: A Practical Guide. Cambridge: Cambridge University Press 2011.

28. Erickson M, Erickson ML, McKnight R, et al. Physical therapy documentation: from examination to outcome: Slack Incorporated 2008.

29. Reiman MP, Manske RC. The assessment of function: How is it measured? A clinical perspective. *Journal of Manual & Manipulative Therapy* 2011;19(2):91-99.

30. Mokkink L. COSMIN risk of bias checklist, 2018.

31. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol* 2009;62(10):e1-34. doi: 10.1016/j.jclinepi.2009.06.006 [published Online First: 2009/07/28]

32. Alamam DM, Leaver A, Moloney N, et al. Pain Behaviour Scale (PaBS): An exploratory study of reliability and construct validity in a chronic low back pain population. *Pain Research and Management* 2019;2019 (no pagination)(2508019)

33. Andersson EI, Lin CC, Smeets R. Performance Tests in People With Chronic Low Back Pain Responsiveness and Minimal Clinically Important Change. *Spine* 2010;35(26):E1559-E63. doi: 10.1097/BRS.0b013e3181cea12e

34. Andrew Walsh D, Jane Kelly S, Sebastian Johnson P, et al. Performance problems of patients with chronic low-back pain and the measurement of patient-centered outcome. *Spine* 2004;29(1):87-93.

35. Armstrong M, McDonough S, Baxter D. Reliability and repeatability of shuttle walk test in patients with chronic low back pain...including commentary by Eiser N, Lemmink KAP, and Walsh DA. *International Journal of Therapy & Rehabilitation* 2005;12(10):438-43.

36. Barz T, Melloh M, Staub L, et al. The diagnostic value of a treadmill test in predicting lumbar spinal stenosis. *European Spine Journal* 2008;17(5):686-90. doi: 10.1007/s00586-008-0593-1

37. Campbell H, Rivero-Arias O, Johnston K, et al. Responsiveness of objective, disease-specific, and generic outcome measures in patients with chronic low back pain: an assessment for improving, stable, and deteriorating patients. *Spine* 2006;31(7):815-22.

38. Caporaso F, Pulkovski N, Sprott H, et al. How well do observed functional limitations explain the variance in Roland Morris scores in patients with chronic non-specific low back pain undergoing physiotherapy? *European Spine Journal* 2012;21:S187-S95. doi: 10.1007/s00586-012-2255-6

39. Conway J, Tomkins CC, Haig AJ. Walking assessment in people with lumbar spinal stenosis: capacity, performance, and self-report measures. *Spine Journal* 2011;11(9):816-23. doi: 10.1016/j.spinee.2010.10.019

40. Cunha IT, Simmonds MJ, Protas EJ, et al. Back pain, physical function, and estimates of aerobic capacity - What are the relationships among methods and measures? *American Journal of Physical Medicine & Rehabilitation* 2002;81(12):913-20. doi: 10.1097/01.Phm.0000030729.77020.2a

41. da Cunha-Filho IT, Lima FC, Guimaraes FR, et al. Use of physical performance tests in a group of Brazilian Portuguese-speaking individuals with low back pain. *Physiotherapy Theory & Practice* 2010;26(1):49-55. doi: 10.3109/09593980802602844

42. Deen HG, Zimmerman RS, Lyons MK, et al. Test-Retest Reproducibility of the Exercise Treadmill Examination in Lumbar Spinal Stenosis. *Mayo Clinic Proceedings* 2000;75(10):1002-07.

43. Denteneer L, van Daele U, Truijen S, et al. Convergent validity of clinical tests which are hypothesized to be associated with physical functioning in patients with nonspecific chronic low back pain. *Journal of Back & Musculoskeletal Rehabilitation* 2019;16:16.

44. Durand MJ, Brassard B, Hong QN, et al. Responsiveness of the Physical Work Performance Evaluation, a functional capacity evaluation, in patients with low back pain. *Journal of Occupational Rehabilitation* 2008;18(1):58-67.

45. Engh L, Strand L, Robinson H, et al. Back Performance Scale: Assessment of patients with back problems in primary health care. *Fysioterapeuten* 2015;82(9):22-7.

46. Felix ZF, Schiltenwolf M, Abel R, et al. Gait analysis does not correlate with clinical and MR imaging parameters in patients with symptomatic lumbar spinal stenosis. *Bmc Musculoskeletal Disorders* 2008;9 doi: 10.1186/1471-2474-9-89

47. Gouttebarge V, Wind H, Kuijer PP, et al. Reliability and Agreement of 5 Ergo-Kit Functional Capacity Evaluation Lifting Tests in Subjects With Low Back Pain. *Archives of Physical Medicine and Rehabilitation* 2006;87(10):1365-70.

48. Grelat M, Gouteron A, Casillas JM, et al. Walking Speed as an Alternative Measure of Functional Status in Patients with Lumbar Spinal Stenosis. *World Neurosurgery* 2019;122:e591-e97.

49. Gulbahar S, Berk H, Pehlivan E, et al. [The relationship between objective and subjective evaluation criteria in lumbar spinal stenosis]. *Acta Orthopaedica et Traumatologica Turcica* 2006;40(2):111-6.

50. Jakobsson M, Brisby H, Gutke A, et al. One-minute stair climbing, 50-foot walk, and timed up-and-go were responsive measures for patients with chronic low back pain undergoing lumbar fusion surgery. *BMC Musculoskeletal Disorders* 2019;20 (1) (no pagination)(137)

51. Kahraman BO, Sengul YS, Kahraman T, et al. Developing a reliable core stability assessment battery for patients with nonspecific low back pain. *Spine* 2016;41(14):E844-E50.

52. Kahraman T, Ozcan Kahraman B, Salik Sengul Y, et al. Assessment of sit-to-stand movement in nonspecific low back pain: a comparison study for psychometric properties of field-based and laboratory-based methods. *International journal of rehabilitation research* 2016;Internationale Zeitschrift fur Rehabilitationsforschung. Revue internationale de recherches de readaptation. 39(2):165-70.

53. Lin SI, Lin RM. Disability and walking capacity in patients with lumbar spinal stenosis: Association with sensorimotor function, balance, and functional performance. *Journal of Orthopaedic & Sports Physical Therapy* 2005;35(4):220-26. doi: 10.2519/jospt.2005.35.4.220

54. Lygren H, Dragesund T, Joensen J, et al. Test-retest reliability of the Progressive Isoinertial Lifting Evaluation (PILE). *Spine (03622436)* 2005;30(9):1070-74.

55. Lüder S, Pfingsten M, Lüdtke K, et al. Kann die Aktivitätskapazität von Patienten mit Rückenschmerzen objektiv und reliabel gemessen werden? *physioscience* 2006;2(04):147-55.

56. Maras G, Citaker S, Meray J. Cross-Cultural Adaptation, Validity, and Reliability Study of the Turkish Version of Back Performance Scale. *Spine* 2019;44(1):E39-E44.

57. Maribo T, Iversen E, Andersen NT, et al. Intra-observer and interobserver reliability of One Leg Stand Test as a measure of postural balance in low back pain patients. *International Musculoskeletal Medicine* 2009;31(4):172-77. doi: 10.1179/175361409X12472218841040

58. Maribo T, Schiottz-Christensen B, Jensen LD, et al. Postural balance in low back pain patients: criterion-related validity of centre of pressure assessed on a portable force platform. *European Spine Journal* 2012;21(3):425-31.

59. Moradi B, Benedetti J, Zahlten-Hinguranage A, et al. The value of physical performance tests for predicting therapy outcome in patients with subacute low back pain: a prospective cohort study. *European Spine Journal* 2009;18(7):1041-9.

60. Ocarino JM, Goncalves GGP, Vaz DV, et al. Correlation between a functional performance questionnaire and physical capability tests among patients with low back pain. *Revista Brasileira De Fisioterapia* 2009;13(4):343-49. doi: 10.1590/s1413-35552009005000046

61. Odebiyi DO, Kujero S, Lawal T. Relationship between spinal mobility, physical performance, pain intensity and functional disability in patients with chronic low back pain. *Nigerian Journal of Medical Rehabilitation* 2006

62. Paatelma M, Karvonen E, Heinonen A. Inter- and intra-tester reliability of selected clinical tests in examining patients with early phase lumbar spine and sacroiliac joint pain and dysfunction. *Advances in Physiotherapy* 2010;12(2):74-80.

63. Pfingsten M, Lueder S, Luedtke K, et al. Significance of physical performance tests for patients with low back pain. *Pain Medicine* 2014;15(7):1211-21.

64. Pozo-Cruz Bd, Triviño-Amigo N, Adsuar-Sala JC, et al. Relative and absolute reliability of the progressive iso-inertial lifting test in patients affected by non-specific, chronic low back pain: a 12-week test-retest study. *Rehabilitacion* 2012;46(4):271-76.

65. Rainville J, Childs LA, Pena EB, et al. Quantification of walking ability in subjects with neurogenic claudication from lumbar spinal stenosis--a comparative study. *Spine Journal: Official Journal of the North American Spine Society* 2012;12(2):101-9.

66. Reneman MF, Jorritsma W, Schellekens JMH, et al. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. *Journal of Occupational Rehabilitation* 2002;12(3):119-29.

67. Ryan CG, Gray H, Newton M, et al. The convergent validity of free-living physical activity monitoring as an outcome measure of functional ability in people with chronic low back pain. *Journal of Back and Musculoskeletal Rehabilitation* 2008;21(2):137-42.

68. Smeets RJEM, Hijdra HJM, Kester ADM, et al. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. *Clinical Rehabilitation* 2006;20(11):989-98.

69. Smith RL. Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation. *Journal of Orthopaedic & Sports Physical Therapy* 1994;19(5):277-81.

70. Staartjes VE, Schroder ML. The five-repetition sit-to-stand test: evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine. *Journal of Neurosurgery Spine* 2018;29(4):380-87.

71. Strand LI, Anderson B, Lygren H, et al. Responsiveness to change of 10 physical tests used for patients with back pain. *Physical Therapy* 2011;91(3):404-15.

72. Taylor S, Frost H, Taylor A, et al. Reliability and responsiveness of the shuttle walking test in patients with chronic low back pain. *Physiotherapy Research International* 2001;6(3):170-8.

73. Tomkins-Lane CC, Battie MC. Validity and reproducibility of self-report measures of walking capacity in lumbar spinal stenosis. *Spine* 2010;35(23):2097-102.

74. Tomkins-Lane CC, Battie MC, Macedo LG. Longitudinal construct validity and responsiveness of measures of walking capacity in individuals with lumbar spinal stenosis. *Spine Journal: Official Journal of the North American Spine Society* 2014;14(9):1936-43.

75. Tomkins CC, Battie MC, Rogers T, et al. A criterion measure of walking capacity in lumbar spinal stenosis and its comparison with a treadmill protocol. *Spine* 2009;34(22):2444-9.

76. Wand BM, Chiffelle LA, O'Connell NE, et al. Self-reported assessment of disability and performance-based assessment of disability are influenced by different patient characteristics in acute low back pain. *European Spine Journal* 2010;19(4):633-40.

77. Wittink H, Rogers W, Sukiennik A, et al. Physical functioning: self-report and performance measures are related but distinct. *Spine* 2003;28(20):2407-13.

78. Reneman MF, Fokkens AS, Dijkstra PU, et al. Testing lifting capacity: validity of determining effort level by means of observation. *Spine* 2005;30(2):E40-6.

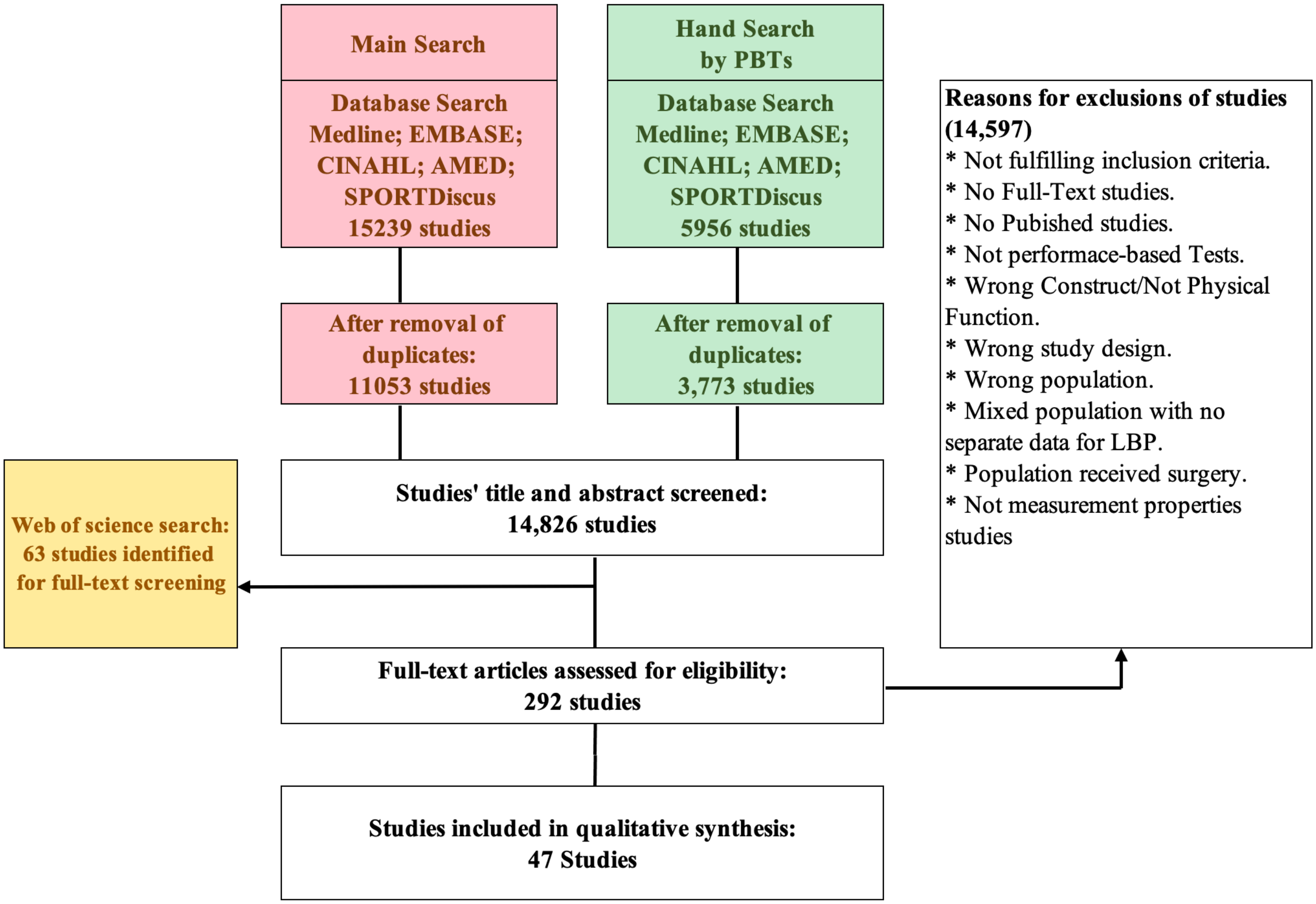
79. Tomkins C, Battie M, Rogers T, et al. A criterion measure of walking capacity in lumbar spinal stenosis and its comparison with a treadmill protocol. *Spine* 2009;34(22):2444-9.

80. Maribo T, Stengaard-Pedersen K, Jensen LD, et al. Postural balance in low back pain patients: Intra-session reliability of center of pressure on a portable force platform and of the one leg stand test. *Gait & Posture* 2011;34(2):213-17.

81. Paatelma M, Karvonen E, Heinonen A. Inter-tester reliability in classifying acute and subacute low back pain patients into clinical subgroups: A comparison of specialists and non-specialists. A pilot study. *Journal of Manual and Manipulative Therapy* 2009;17(4):221-29.

3.6 Figures

Figure 1: the flow diagram of studies' search and selection.



3.7 Results Tables

3.7.1 Table 3: Characteristics of included studies.

No.	Authors	Sample Size	Back Pain Type	Back Pain Duration	Eligible PBMs	Psychometric property evaluated	COSMIN ROB
1	Alaman, D. M., et al. (2019)	22	Non-Specific LBP	Chronic more than 3 months	5-Repeated Sit To Stand (s) Timed Up And Go (s) 50-Foot Walk Test (s)	Reliability: Test-Retest, Inter-Rater, and Intra-Rater	Doubtful
2	Andersson, E. I., et al. (2010)	198	Non-Specific LBP	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) 5-Repeated Sit To Stand (s) Stair Climbing (steps) Progressive Isoinertial Lifting Evaluation (cycles)	Responsiveness (Criterion Approach)	Inadequate
3	Andrew Walsh, D., et al. (2004)	101	Mixed LBP population	Chronic at least 12 months	5-Minute Walk Test (m)	Responsiveness (Construct Approach)	Inadequate
4	Armstrong, M., et al. (2005)	10	Mixed LBP population	Chronic at least 6 months	Shuttle Walk Test (m)	Test-Retest Reliability	Doubtful
5	Barz, T., et al. (2008)	25	Lumber Spinal Stenosis	Not reported: Assumed Chronic	Ambulatory-Treadmill test (m)	Convergent Validity	Inadequate
6	Campbell, H., et al. (2006)	250	Mixed LBP population	Chronic at least 12 months	Shuttle Walking Test (m)	Responsiveness (Criterion Approach)	Very Good
7	Caporaso, F., et al. (2012)	37	Non-Specific LBP	Chronic more than 3 months	Sock Test Sit-Up Test Stand To floor Lift Test Stair Climb Pick-Up Test Functional Test Index	Convergent Validity Responsiveness (Construct Approach)	Adequate
8	Conway, J., et al. (2011)	12	Lumber Spinal Stenosis	Chronic (average of 4 years)	Self-Paced Walking Test (m)	Convergent Validity	Doubtful

9	Cunha, I. T., et al. (2002)	51	Mixed LBP population	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) 5-Repeated Sit To Stand (s) Roll from right to left (s)	Convergent Validity	Inadequate
10	da Cunha-Filho, I. T., et al. (2010)	30	Mixed LBP population	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) 5-Repeated Sit To Stand (s) Timed Up And Go (s)	Convergent Validity Test-Retest Reliability	Doubtful
11	Deen, H. G., et al. (2000)	28	Lumber Spinal Stenosis	Chronic more than 3 months	Treadmill Examination (min)	Test-Retest Reliability	Inadequate
12	Denteneer, L., et al. (2019)	25	Non-Specific LBP	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s)	Convergent Validity (only btw PBMs)	Doubtful
13	Durand, M. J., et al. (2008)	27	Non-Specific LBP	Chronic more than 6 months	Physical Work Performance Evaluation	Responsiveness (Construct Approach)	Inadequate
14	Engh, L., et al. (2015)	52	Mixed LBP population	64 % Chronic 11 % Acute/subacute	Back Performance Scale	Convergent Validity Test-Retest Reliability known-groups validity	Doubtful
15	Felix, Z. F., et al. (2008)	63	Lumber Spinal Stenosis	Chronic at least 6 months	Treadmill Walking Test (m)	Convergent Validity	Inadequate
16	Gouttebarge, V., et al. (2006)	24	Lumber Spinal Stenosis	Chronic more than 12 months	Back-Torso Lift Test (kg) Shoulder Lift Test (kg) Carrying Lifting Strength Test (kg) Lower Lifting Strength Test (kg) Upper Lifting Strength Test (kg)	Inter-Rater Reliability	Very Good
17	Grelat, M., et al. (2019)	38	Lumber Spinal Stenosis	Chronic more than 3 months	6-Minute Walk Test (m) Free Walking Velocity Test (m/s)	Convergent Validity	Inadequate
18	Gulbahar, S., et al. (2006)	30	Lumber Spinal Stenosis	Chronic more than 3 months	Treadmill Tolerance Test (s)	Convergent Validity	Inadequate
19	Jakobsson, M., et al. (2019)	118	Degenerative LBP	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) Stair Climbing (steps) Timed Up And Go (s)	Responsiveness (Construct and Criterion Approach)	Very Good

20	Kahraman, B. O., et al. (2016)	38	Non-Specific LBP	Chronic more than 3 months	30-sec Right and Left Lateral Step-Down (no. of repetition)	Intra-Rater Reliability	Doubtful
21	Kahraman, T., et al. (2016)	38	Non-Specific LBP	Chronic more than 3 months	30-sec Chair Stand Test (no. of repetition)	Intra-Rater Reliability	Doubtful
22	Lin, S. and R. Lin (2005)	50	Lumber Spinal Stenosis	Chronic more than 12 months	Single-Leg Stance Test (s) Timed Up And Go (s)	Convergent Validity	Inadequate
23	Lygren, H., et al. (2005)	31	Muscular-related Low Back Pain	Chronic more than 12 months	Progressive Isonertial Lifting Evaluation	Test-Retest Reliability	Doubtful
24	Lueder, S., et al. (2006)	59	Mixed LBP population	Not reported: Assumed Chronic	Stair Climbing Pick-Up Test Rising–Up Test Lacing Test Sock Test Sit-Up Test Hair-Wash Test Lift Test Stand-To-Floor Test	Reliability (Inter-Rater, and Intra-Rater)	Doubtful
25	Maras, G., et al. (2019)	180	Mixed LBP population	Chronic more than 3 months	Back Performance Scale Tests (BPS): BPS-Sock Test BPS-Pick Up Test BPS-Roll Up BPS-Lifting Test Back Performance Scale-total	Test-Retest Reliability	Doubtful
					Back Performance Scale- total	Convergent Validity	Very Good
26	Maribo, T., et al. (2009)	48	Mixed LBP population	Chronic more than 6 months	Single-Leg Stance Test (s)	Inter-rater reliability and Intra-rater reliability	Inadequate
27	Maribo, T., et al. (2011)	52	Mixed LBP population	Chronic at least 3 months	Single-Leg Stance Test (s)	Intra-Rater Reliability	Doubtful
28	Moradi, B., et al. (2009)	162	Mixed LBP population	Chronic more than 12 months	Villager Test	Convergent Validity	Inadequate
29	Ocarino, J. M., et al. (2009)	30	Non-Specific LBP	Chronic more than 3 months	50-Foot Walk Test (s) 5-Repeated Sit To Stand (s)	Convergent Validity	Inadequate

30	Odebiyi, D. O., et al. (2006)	23	Mixed LBP population	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) 5-Repeated Sit To Stand (s) 360-deg Roll Over (s)	Convergent Validity	Inadequate
31	Paatelma, M., et al. (2010)	15	Mixed LBP population	27 % Acute 73 % Subacute	Single-Leg Stance Test (s) Functional Battery Test	Inter-rater reliability and Intra-rater reliability	Doubtful
32	Pfingsten, M., et al. (2014)	106	Non-Specific LBP	Chronic more than 3 months	Stair Climbing Stand To Floor Lift Test Sock Test Roll-Up Test Pick-Up Test	Convergent Validity (only btw PBMs)	Inadequate
						Known-Group Validity	Very Good
33	Pozo-Cruz, B. d., et al. (2012)	10	Non-Specific LBP	Chronic at least 6 months	Progressive Isonertial Lifting Evaluation	Test-Retest Reliability	Doubtful
34	Rainville, J., et al. (2012)	50	Lumber Spinal Stenosis	Chronic more than 12 months	Self-Paced Walk Test (m) Motorized Treadmill Test (s)	Responsiveness (Construct Approach)	Inadequate
						Responsiveness (Criterion Approach)	Adequate
35	Reneman, M. F., et al. (2002)	64	Non-Specific LBP	Chronic more than 6 months	Isernhagen Work Systems Functional Capacity Evaluation	Convergent Validity	Very Good
					Shuttle Walk Test (m)	Convergent Validity	Doubtful
36	Reneman, M. F., et al. (2005)	16	Non-Specific LBP	Chronic more than 3 months	Isernhagen Work Systems Functional Capacity Evaluation	Inter-Rater Reliability	Very Good
37	Ryan, C. G., et al. (2008).	38	Non-Specific LBP	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) 5-Repeated Sit To Stand (s)	Convergent Validity	Inadequate
38	Smeets, R. J. E. M., et al. (2006)	53	Non-Specific LBP	Chronic more than 3 months	5-Minute Walk Test (m) 50-Foot Walk Test (s) 5-Repeated Sit To Stand (s) Stair Climbing Progressive Isonertial Lifting Evaluation	Test-Retest Reliability	Doubtful

39	Smith, R. L. (1994)	21	Degenerative LBP	Chronic more than 12 months	Functional Capacity Evaluation (safe maximum lifting)	Inter-rater reliability and Intra-rater reliability	Adequate
40	Staartjes, V. E. and M. L. Schroder (2018)	150	Degenerative LBP	No pain = 8 6 wks-6 mons = 27 6 mons -1 yr.= 42 > 1 yr.= 80	5-Repeated Sit To Stand (s)	Convergent Validity	Inadequate
41	Strand, L. I., et al. (2011)	98	Non-Specific LBP	Assumed Chronic	Lift Test 15-Meter Walk Test Back Performance Scale Progressive Isonertial Lifting Evaluation	Test-Retest Reliability	Inadequate
						Convergent Validity Responsiveness (Construct Approach)	Very Good
42	Taylor, S., et al. (2001)	44	Non-Specific LBP	Chronic at least 6 months	Shuttle Walk Test (m)	Test-Retest Reliability	Doubtful
43	Tomkins-Lane, C. C. and M. C. Battie (2010)	49	Lumber Spinal Stenosis	Chronic more than 12 months	Self-Paced Walking Test (m)	Convergent Validity	Very Good
44	Tomkins-Lane, C. C., et al. (2014)	26	Lumber Spinal Stenosis	Chronic more than 12 months	Self-Paced Walking Test (m)	Responsiveness (Construct and Criterion Approach)	Doubtful
45	Tomkins, C. C., et al. (2009)	45	Lumber Spinal Stenosis	Chronic more than 12 months	Self-Paced Walking Test (total distance and time)	Test-Retest Reliability	Doubtful
					Treadmill Walk Test (m)	Convergent Validity	Very Good
46	Wand, B. M., et al. (2010)	94	Non-Specific LBP	Acute less than 6 weeks	Timed Functional Tests Total Score Timed Sit To Stand Test Timed Up and Go Timed 5-Minute Walk Test Timed Lying To Stand Test	Convergent Validity (only btw PBMs)	Inadequate
47	Wittink, H., et al. (2003)	75	Mixed LBP population	Chronic more than 3 months	Modified-Symptom-Limited Treadmill Test (time)	Convergent Validity	Inadequate

3.7.2 Table 4: A summary of Test-Retest Reliability results

Performance Based Measurements	Studies	Back Pain Type	ICC/Kappa Scores (95% CI)	ICC or Kappa ≥ 0.70	COSMIN ROB	PBM Overall Result Rating	Level of Evidence
5-Min Walk Test (m)	Smeets, R. J. E. M., et al. (2006)	Non-Specific LBP	0.89 (0.81-0.93)	+	Doubtful	+	Limited
	da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	0.94	+	Doubtful	+	Limited
50-ft Walk Test (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	1.0 (0.9-1.0)	+	Doubtful	+	Moderate
	Smeets, R. J. E. M., et al. (2006)	Non-Specific LBP	0.76 (0.61-0.85)*transformed-inverse	+	Doubtful		
	Strand, L. I., et al. (2011)	Non-Specific LBP	0.77 (0.24-0.94)	+	Inadequate		
	da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	0.94	+	Doubtful	+	Limited
Time Up-and-Go (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.8 (0.6-0.9)	+	Doubtful	+	Limited
	da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	0.98	+	Doubtful	+	Limited
5-repetition sit-to-stand test (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.6 (0.3-0.8)	–	Doubtful	–	Limited
	Smeets, R. J. E. M., et al. (2006)	Non-Specific LBP	0.91 (0.81-0.94) *transformed - inverse	+	Doubtful		
	da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	0.99	+	Doubtful	+	Limited
	Staartjes, V. E. and M. L. Schroder (2018)	Degenerative LBP	0.97 (0.94-0.98)	+	Inadequate	+	Poor
Stair Climbing (steps)	Smeets, R. J. E. M., et al. (2006)	Non-Specific LBP	0.96 (0.93-0.98)	+	Doubtful	+	Limited

Shuttle Walk Test (m)		Armstrong, M., et al. (2005)	Mixed LBP population	Pearson's r=0.98	?	Doubtful	?	Limited
		Taylor, S., et al. (2001)	Non-Specific LBP	0.99	+	Doubtful	+	Limited
Self-Paced Walk Test	Total ambulation distance	Tomkins, C. C., et al. (2009)	Lumber Spinal Stenosis	0.98 (0.95-0.99)	+	Doubtful	+	Limited
	Distance-first symptom	Tomkins, C. C., et al. (2009)	Lumber Spinal Stenosis	0.94 (0.89-0.97)	+	Doubtful	+	Limited
	Speed	Tomkins, C. C., et al. (2009)	Lumber Spinal Stenosis	0.80 (0.64-0.90)	+	Doubtful	+	Limited
Treadmill Walk Test	Treadmill 1.2 speed (time to first symptom)	Deen, H. G., et al. (2000)	Lumber Spinal Stenosis	0.90	+	Inadequate	+	Poor
	Treadmill 1.2 speed (total ambulation time)	Deen, H. G., et al. (2000)	Lumber Spinal Stenosis	0.89	+	Inadequate	+	Poor
	Treadmill preferred speed (time to first symptom)	Deen, H. G., et al. (2000)	Lumber Spinal Stenosis	0.98	+	Inadequate	+	Poor
	Treadmill preferred speed (total ambulation time)	Deen, H. G., et al. (2000)	Lumber Spinal Stenosis	0.96	+	Inadequate	+	Poor
BPS-sock test		Engh, L., et al. (2015)	Mixed LBP population	0.65 (0.48-0.81)	–	Doubtful	–	Limited
		Maras, G., et al. (2019)	Mixed LBP population	0.897 (0.830-0.937)	+	Doubtful		
BPS-Pick up test		Engh, L., et al. (2015)	Mixed LBP population	0.53 (0.29-0.78)	–	Doubtful	–	Limited
		Maras, G., et al. (2019)	Mixed LBP population	0.857 (0.766-0.913)	+	Doubtful		
BPS-Roll up		Engh, L., et al. (2015)	Mixed LBP population	0.53 (0.33-0.73)	–	Doubtful	–	Limited
		Maras, G., et al. (2019)	Mixed LBP population	0.899 (0.835-0.939)	+	Doubtful		

BPS-Lifting test		Engh, L., et al. (2015)	Mixed LBP population	0.57 (0.35-0.80)	–	Doubtful	–	Limited
		Maras, G., et al. (2019)	Mixed LBP population	0.795 (0.664-0.875)	+	Doubtful		
Lifting Tests (no. of lifts in 1 min)		Strand, L. I., et al. (2011)	Non-Specific LBP	0.87 (0.50-0.97)	+	Inadequate	+	Poor
Progressive Isoinertial Lifting Evaluation (PILE)	PILE (highest load, kg)	Lygren, H., et al. (2005)	Muscular-related Low Back Pain	0.91 (– 4.5-4.5)	+	Doubtful	+	Limited
		Pozo-Cruz, B. d., et al. (2012)	Non-Specific LBP	0.96 (0.88-0.98)	+	Doubtful	+	Limited
		Strand, L. I., et al. (2011)	Non-Specific LBP	0.91 (0.65, 0.98)	+	Inadequate		
	PILE (lifting stages)	Smeets, R. J. E. M., et al. (2006)	Non-Specific LBP	0.92 (0.87-0.96)	+	Doubtful	+	Limited
Back Performance Scale (0-15 points)		Maras, G., et al. (2019)	Mixed LBP population	0.905 (0.867-0.936)	+	Doubtful	+	Moderate
		Engh, L., et al. (2015)	Mixed LBP population	0.93 (0.87-0.96)	+	Doubtful		
		Strand, L. I., et al. (2011)	Non-Specific LBP	0.89 (0.51-0.98)	+	Inadequate	+	Poor

BPS; Back Performance Scale

3.7.3 Table 5: A summary of Intra-Rater Reliability results

Performance Based Tests	**Studies**	**Back Pain Type**	**ICC/Kappa Scores (95% CI)**	**ICC or Kappa ≥ 0.70**	**COSMIN ROB**	**PBM Overall Result Rating**	**Level of Evidence**
Repeated Sit to Stand (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.8 (0.5-0.9)	+	Doubtful	+	Limited
30-s chair stand test	Kahraman, T., et al. (2016)	Non-Specific LBP	0.94 (0.89-0.97)	+	Doubtful	+	Limited
Timed up and go (TUG) (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.9 (0.8-1.0)	+	Doubtful	+	Limited
50-foot walk (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.8 (0.6-0.9)	+	Doubtful	+	Limited

Stair Climbing	Lüder S., et al. (2006)	Mixed LBP population	0.59 (70%-11%)	–	Doubtful	–	Limited
Lateral Step-down	Kahraman, B. O., et al. (2016)	Non-Specific LBP	R: 0.93 (0.87-0.96) L: 0.92 (0.86- 0.96)	+	Doubtful	+	Limited
Pick-up test	Lüder S., et al. (2006)	Mixed LBP population	0.69 (80%-7%)	–	Doubtful	–	Limited
Rising–Up Test	Lüder S., et al. (2006)	Mixed LBP population	0.66 (78%-5%)	–	Doubtful	–	Limited
Lacing Test	Lüder S., et al. (2006)	Mixed LBP population	0.84 (89%-2%)	+	Doubtful	+	Limited
Sock test	Lüder S., et al. (2006)	Mixed LBP population	0.88 (93%-4%)	+	Doubtful	+	Limited
Sit-Up test	Lüder S., et al. (2006)	Mixed LBP population	0.59% (72%-0%)	–	Doubtful	–	Limited
Hair-Wash Test	Lüder S., et al. (2006)	Mixed LBP population	0.77 (83%-4%)	+	Doubtful	+	Limited
Stand-to-Floor Test	Lüder S., et al. (2006)	Mixed LBP population	0.57 (71%-10%)	–	Doubtful	–	Limited
Lift test	Lüder S., et al. (2006)	Mixed LBP population	0.79 (85%-4%)	+	Doubtful	+	Limited
FCE-safe maximum lifting	Smith, R. L. (1994)	Degenerative LBP	0.73	+	Adequate	+	Moderate
One Leg Stand Test	Maribo, T., et al. (2009)	Mixed LBP population	0.86	+	Inadequate	+	Moderate
	Maribo, T., et al. (2011)	Mixed LBP population	0.79	+	Doubtful		
	Paatelma, M., et al. (2010)	Mixed LBP population	0.59 (0.04-0.89)	–	Doubtful		
Single-legged hop test	Kahraman, B. O., et al. (2016)	Non-Specific LBP	R: 0.98 (0.96-0.99) L: 0.97 (0.94-0.98)	+	Doubtful	+	Limited
Function Batter Test	Paatelma, M., et al. (2010)	Mixed LBP population	0.9 (0.4-1.2)	+	Doubtful	+	Limited

3.7.4 Table 6: A summary of Inter-Rater Reliability results

Performance Based Tests	Studies	Back Pain Type	ICC/Kappa Scores (95% CI)	ICC or Kappa ≥ 0.70	COSMIN ROB	PBM Overall Result Rating	Level of Evidence
Repeated Sit to Stand (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	1.0 (0.9-1.0)	+	Doubtful	+	Limited
Timed up and go (TUG) (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.9 (0.9-1.0)	+	Doubtful	+	Limited
50-foot walk (s)	Alamam, D. M., et al. (2019)	Non-Specific LBP	0.9 (0.8-1.0)	+	Doubtful	+	Limited
Stair Climbing	Lüder S., et al. (2006)	Mixed LBP population	0.33 (52%-5%)	–	Doubtful	–	Limited
Pick-up test	Lüder S., et al. (2006)	Mixed LBP population	0.48 (69%-15%)	–	Doubtful	–	Limited
Rising–Up Test	Lüder S., et al. (2006)	Mixed LBP population	0.44 (64%-16%)	–	Doubtful	–	Limited
Lacing Test	Lüder S., et al. (2006)	Mixed LBP population	0.61 (72%-8%)	–	Doubtful	–	Limited
Sock test	Lüder S., et al. (2006)	Mixed LBP population	0.56 (71%-11%)	–	Doubtful	–	Limited
Sit-Up test	Lüder S., et al. (2006)	Mixed LBP population	0.46 (62%-4%)	–	Doubtful	–	Limited
Hair-Wash Test	Lüder S., et al. (2006)	Mixed LBP population	0.16 (33%-33%)	–	Doubtful	–	Limited
Lift test	Lüder S., et al. (2006)	Mixed LBP population	0.45 (61%-11%)	–	Doubtful	–	Limited
Stand-to-Floor Test	Lüder S., et al. (2006)	Mixed LBP population	0.42 (60%-17%)	–	Doubtful	–	Limited
One Leg Stand Test	Maribo, T., et al. (2009)	Mixed LBP population	1.42 (1.12-1.95)	+	Inadequate	–	Limited
	Paatelma, M., et al. (2010)	Mixed LBP population	0.67 (0.32-1.00)	–	Doubtful		
FCE-safe maximum lifting	Smith, R. L. (1994)	Degenerative LBP	Rater 1: 0.64 Rater 2: 0.62	–	Adequate	–	Limited
IWS FCE- Lifting Test Borg CR-10 scale	Reneman, M. F., et al. (2005)	Non-Specific LBP	0.76 (0.69-0.83)	+	Very Good	+	Strong
IWS FCE- Lifting Test Categorical Scale	Reneman, M. F., et al. (2005)	Non-Specific LBP	0.5	–	Very Good	–	Limited

Back-torso lift test	Gouttebarge, V., et al. (2006)	Lumber Spinal Stenosis	0.97 (0.94-0.99)	+	Very Good	+	Strong
Shoulder lift test	Gouttebarge, V., et al. (2006)	Lumber Spinal Stenosis	0.96 (0.91-0.98)	+	Very Good	+	Strong
Carrying lifting strength test	Gouttebarge, V., et al. (2006)	Lumber Spinal Stenosis	0.95 (0.84-0.98)	+	Very Good	+	Strong
Lower lifting strength test	Gouttebarge, V., et al. (2006)	Lumber Spinal Stenosis	0.94 (0.85-0.97)	+	Very Good	+	Strong
Upper lifting strength test	Gouttebarge, V., et al. (2006)	Lumber Spinal Stenosis	0.95 (0.89-0.98)	+	Very Good	+	Strong
Function Battery Test	Paatelma, M., et al. (2010)	Mixed LBP population	0.9 (0.4-1.2)	+	Doubtful	+	Limited

3.7.5 Table 7: Summary of results for PBMs' convergent validity.

Performance Based Measurements PBMs	PROM	Studies	Back Pain Type	Correlation Cut-off	Correlation Scores	Results Ratings	COSMIN ROB	PBM Overall Result Rating	Level of Evidence
5-Minute Walk Test	**RMDQ**	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	–0.41	–	Very Good	–	Limited
		da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	related constructs ≥ 0.50	–0.39	–	Very Good		
		Odebiyi, D. O., et al (2006)	Mixed LBP population	related constructs ≥ 0.50	0.253	–	Doubtful		
		Ryan, C. G., et al. (2008)	Non-Specific LBP	related constructs ≥ 0.50	–0.25	–	Very Good	–	Limited
	COPM walk (performance)	Andrew Walsh, D., et al. (2004)	Mixed LBP population	related constructs ≥ 0.50	0.27	–	Doubtful	–	Limited
	COPM walk (satisfaction)	Andrew Walsh, D., et al. (2004)	Mixed LBP population	related constructs ≥ 0.50	–0.02	–	Doubtful	–	Limited

	Self-Reported PA	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	0.33	–	Very Good	–	Limited
Shuttle WT (distance)	**ODI-walking**	Campbell, H., et al. (2006)	Mixed LBP population	related constructs ≥ 0.50	–0.62	+	Very Good	+	Strong
		Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	–0.17	–	Doubtful	–	Limited
	EQ-5D (Q 1)	Campbell, H., et al. (2006)	Mixed LBP population	related constructs ≥ 0.50	–0.45	–	Doubtful	–	Limited
	SF-36 (Q3.7)	Campbell, H., et al. (2006)	Mixed LBP population	related constructs ≥ 0.50	0.45	–	Doubtful	–	Limited
	SF-36 (Q3.8)	Campbell, H., et al. (2006)	Mixed LBP population	related constructs ≥ 0.50	0.62	+	Doubtful	+	Limited
	SF-36 (Q3.9)	Campbell, H., et al. (2006)	Mixed LBP population	related constructs ≥ 0.50	0.56	+	Doubtful	+	Limited
	RMDQ-17	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	Somers' d index *d* = 0.03	?	Doubtful	?	Limited
	Quebec-8	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	– 0.27	–	Doubtful	–	Limited
	Quebec-9	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	–0.32	–	Doubtful	–	Limited
Shuttle WT (Speed)	**RMDQ-3**	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	Somers' d index *d* = 0.13	–	Doubtful	–	Limited
50-ft Walk Test (s) Time	**RMDQ**	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	0.44	–	Very Good	–	Limited
		da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	related constructs ≥ 0.50	0.23	–	Very Good		

		Odebiyi, D. O., et al (2006)	Mixed LBP population	related constructs ≥ 0.50	0.456	–	Doubtful		
		Ryan, C. G., et al. (2008)	Non-Specific LBP	related constructs ≥ 0.50	0.23	–	Very Good	–	Limited
		Ocarino, J. M., et al. (2009)	Non-Specific LBP	related constructs ≥ 0.50	0.2481	–	Doubtful		
	Self-Reported PA	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	–0.18	–	Very Good	–	Limited
50-ft walk test (m/s) speed	**RMDQ**	Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	–0.37	–	Very Good	–	Limited
	FFbH-R	Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	–0.40	–	Very Good	–	Limited
Ambulatory-Treadmill test (distance)	**ODI**	Barz, T., et al. (2008)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.51	+	Very Good	+	Strong
	Patient Expectations (walk distance)	Barz, T., et al. (2008)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.62	+	Very Good	+	Strong
Modified-Symptom-Limited Treadmill Test (time)	**SF-36 (physical functioning domain)**	Wittink, H., et al. (2003)	Mixed LBP population	related constructs ≥ 0.50	0.43	–	Very Good	–	Limited
Motorised Treadmill Test (Time)	**Estimated walking time**	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.73	+	Adequate	+	Moderate
	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.66	+	Adequate	+	Moderate
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.63	+	Adequate	+	Moderate

	SSQ Physical Function	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.63	+	Adequate	+	Moderate
Motorised Treadmill Test (Distance)	**Estimated walking time**	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.7	+	Adequate	+	Moderate
	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.72	+	Adequate	+	Moderate
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.54	+	Adequate	+	Moderate
	SSQ Physical Function	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.45	–	Adequate	–	Limited
Treadmill walking test	**subjective estimation of walking distance**	Felix, Z. F., et al. (2008)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.121	–	Doubtful	–	Limited
Treadmill Tolerance Test (Time)	**ODI**	Gulbahar, S., et al. (2006)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.54	+	Very Good	+	Strong
	SF-36 (physical functioning domain)	Gulbahar, S., et al. (2006)	Lumber Spinal Stenosis	unrelated constructs ≥ 0.3	0.51	+	Very Good	+	Strong
Timed up and go (s)	**ODI**	Lin, S. and R. Lin (2005)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.446	–	Very Good	–	Limited
	PFS	Lin, S. and R. Lin (2005)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.530	+	Very Good	+	Strong
	RMDQ	da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	related constructs ≥ 0.50	0.17	–	Very Good	–	Limited

6-Minute Walk Test	**ODI**	Grelat, M., et al. (2019)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.44	–	Adequate	–	Limited
	Quebec	Grelat, M., et al. (2019)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.31	–	Adequate	–	Limited
Free walking velocity test	**ODI**	Grelat, M., et al. (2019)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.51	+	Adequate	+	Moderate
	Quebec	Grelat, M., et al. (2019)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.51	+	Adequate	+	Moderate
Self-Paced Walking Test (Distance)	**Quebec**	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.638	+	Very Good	+	Strong
	SF-36 (physical functioning domain)	Conway, J., et al. (2011)	Lumber Spinal Stenosis	unrelated constructs ≥ 0.3	0.825	+	Very Good	+	Strong
	ODI	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.595	+	Very Good	+	Strong
		Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.52	+	Very Good		
	ODI-walking	Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.83	+	Very Good	–	Limited
		Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.49	–	Adequate		
	SSQ Physical Function	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.610	+	Very Good	+	Strong

		Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.62	+	Very Good		
		Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.55	+	Adequate		
	Swiss.PF_Walk Item	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.715	+	Very Good	+	Strong
		Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.66	+	Very Good		
	Self-estimated Walking Distance	Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.8	+	Very Good	+	Strong
		Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.65	+	Adequate		
		Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.886	+	Very Good		
	self-estimated Walk Cap (ordinal 0-10	Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.65	+	Very Good	+	Strong
		Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.682	+	Very Good		
	Self-Estimated walking Time	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.63	+	Adequate	+	Moderate
	HUI3 Amb	Tomkins-Lane, C. C. and M. C. Battie (2010)	Lumber Spinal Stenosis	unrelated constructs ≥ 0.3	0.71	+	Very Good	+	Strong

	Swiss.SS_Weak	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.742	+	Very Good	+	Strong
	Swiss.SS_Balance	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.673	+	Very Good	+	Strong
	Leg pain	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.492	–	Very Good	–	Limited
	Quebec_Stand	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.551	+	Very Good	+	Strong
	Quebec_Walk	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.755	+	Very Good	+	Strong
	Quebec_Reach	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.650	+	Very Good	+	Strong
	Quebec_Run	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.664	+	Very Good	+	Strong
	Quebec_Groceries	Conway, J., et al. (2011)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.727	+	Very Good	+	Strong
Self-Paced Walking Test (Time)	**Estimated walking time**	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.56	+	Adequate	+	Moderate
	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.58	+	Adequate	+	Moderate
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.47	–	Adequate	–	Limited
	SSQ Physical Function	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.58	+	Adequate	+	Moderate

5-repetition Sit to Stand (Time)	**RMDQ**	Ryan, C. G., et al. (2008)	Non-Specific LBP	related constructs ≥ 0.50	0.436	–	Very Good	–	Limited
		Ocarino, J. M., et al. (2009)	Non-Specific LBP	related constructs ≥ 0.50	0.38	–	Doubtful		
		Odebiyi, D. O., et al (2006)	Mixed LBP population	related constructs ≥ 0.50	0.178	–	Doubtful	–	Limited
		Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	0.48	–	Very Good		
		da Cunha-Filho, I. T., et al. (2010)	Mixed LBP population	related constructs ≥ 0.50	0.44	–	Very Good		
		Staartjes, V. E. and M. L. Schroder (2018)	Degenerative LBP	related constructs ≥ 0.50	0.49	–	Very Good	–	Limited
	ODI	Staartjes, V. E. and M. L. Schroder (2018)	Degenerative LBP	related constructs ≥ 0.50	0.44	–	Very Good	–	Limited
	EQ-5D index	Staartjes, V. E. and M. L. Schroder (2018)	Degenerative LBP	Unrelated constructs ≥ 0.3	–0.41	+	Very Good	+	Strong
	Self-Reported PA	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	–0.20	–	Very Good	–	Limited
30s-chair stand test	**ODI**	Kahraman, T., et al. (2016) Assessment	Non-Specific LBP	related constructs ≥ 0.50	–0.442	–	Very Good	–	Limited
20 Steps–Stair Climbing (Time)	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	baseline: 0.49 Post-Treat.: 0.52	–	Very Good	–	Limited
Rolling R and L Tests (s)	**RMDQ**	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	0.44	–	Very Good	–	Limited

	Self-Reported PA	Cunha, I. T., et al. (2002)	Mixed LBP population	related constructs ≥ 0.50	–0.10	–	Very Good	–	Limited
360 Roll-Over Test	**RMDQ**	Odebiyi, D. O., et al (2006)	Mixed LBP population	related constructs ≥ 0.50	0.31	–	Doubtful	–	Limited
Back Performance Scale	**FFbH-R**	Engh, L., et al. (2015)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.68	+	Very Good	+	Strong
		Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	0.56	+	Very Good	+	Strong
	RMDQ	Maras, G., et al. (2019)	Mixed LBP population	related constructs ≥ 0.50	0.576	+	Doubtful	+	Limited
		Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	0.44	–	Very Good	–	Limited
	ODI	Maras, G., et al. (2019)	Mixed LBP population	related constructs ≥ 0.50	0.603	+	Doubtful	+	Limited
Progressive Isoinertial Lifting Evaluation	**FFbH-R**	Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	–0.44	–	Very Good	–	Limited
	RMDQ	Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	–0.32	–	Very Good	–	Limited
Functional Capacity Evaluation	**RMDQ**	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	–0.20	–	Very Good	–	Limited
	ODI	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	–0.52	+	Very Good	+	Strong
	Quebec	Reneman, M. F., et al. (2002) validity	Non-Specific LBP	related constructs ≥ 0.50	–0.50	+	Very Good	+	Strong

Single Leg Stance Tests	**ODI**	Lin, S. and R. Lin (2005)	Lumber Spinal Stenosis	related constructs ≥ 0.50	–0.225	–	Very Good	–	Limited
	PFS	Lin, S. and R. Lin (2005)	Lumber Spinal Stenosis	related constructs ≥ 0.50	0.082	–	Very Good	–	Limited
Stand to Floor Tests	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	baseline: 0.40 Post-Treat.: 0.51	–	Very Good	–	Limited
Sock Tests	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	baseline: 0.27 Post-Treat.: 0.33	–	Very Good	–	Limited
Pick-up test	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	baseline: 0.56 Post-Treat.: 0.15	–	Very Good	–	Limited
5-Repetition Lift Test (ordinal 0-4)	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	**baseline: 0.49** Post-Treat.: 0.52	–	Very Good	–	Limited
Lift test (no. of lifts in 1 min)	**FFbH-R**	Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	–0.42	–	Very Good	–	Limited
	RMDQ	Strand, L. I., et al. (2011)	Non-Specific LBP	related constructs ≥ 0.50	–0.38	–	Very Good	–	Limited
Functional Test Index	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	**baseline: 0.60** Post-Treat.: 0.70	+	Very Good	+	Strong
Sit-Up Test	**RMDQ**	Caporaso, F., et al. (2012)	Non-Specific LBP	related constructs ≥ 0.50	**baseline: 0.36** Post-Treat.: 0.48	–	Very Good	–	Limited

Villiger Test (No. of steps) baseline	**FFbH-R**	Moradi, B., et al. (2009)	Mixed LBP population	related constructs ≥ 0.50	–0.40	–	Doubtful	–	Limited
	PDI	Moradi, B., et al. (2009)	Mixed LBP population	related constructs ≥ 0.50	–0.40	–	Doubtful	–	Limited
Villiger Test (Test Duration) Baseline	**FFbH-R**	Moradi, B., et al. (2009)	Mixed LBP population	related constructs ≥ 0.50	–0.36	–	Doubtful	–	Limited
	PDI	Moradi, B., et al. (2009)	Mixed LBP population	related constructs ≥ 0.50	–0.35	–	Doubtful	–	Limited

COPM walk (performance), Reported Performance With Walk Tolerance from Canadian Occupational Performance Measure; COPM walk (satisfaction), Reported Satisfaction With Walk Tolerance from Canadian Occupational Performance Measure; EQ-5D (Q 1), Mobility Item from Euro-Quality Of Life 5 Domains; EQ-5D index, Euro-Quality Of Life 5 Domains; FFbH-R, Hannover Functional Ability Questionnaire for Measuring Back Pain-Related Disability; HUI3 Amb, Health Utilities Index Single Attribute Utility Score for Ambulation; ODI, Oswestry Disability Index; ODI-walking, Walking Distance Item from The Oswestry Disability Index; PDI, Pain Disability Index; PFS, Physical Functional Scale; Quebec, Quebec Back Pain Disability Scale; Quebec_Groceries, Groceries Item from Quebec Back Pain Disability Scale; Quebec_Reach, Reach Item from Quebec Back Pain Disability Scale; Quebec_Run, Run Item from Quebec Back Pain Disability Scale; Quebec_Stand, Stand Item from Quebec Back Pain Disability Scale; Quebec_Walk, Walk Item from Quebec Back Pain Disability Scale; Quebec-8, Question 8 from Quebec Back Pain Disability Scale; Quebec-9, Question 9 from Quebec Back Pain Disability Scale; RMDQ, Roland-Morris Disability Questionnaire; RMDQ-17, Question 17 from Roland-Morris Disability Questionnaire; RMDQ-3, Question 3 from Roland-Morris Disability Questionnaire; SF-36 (physical functioning domain), Physical Functioning Domain from 36-Item Short Form Health Survey; SF-36 (Q3.7), Question 3.7 from 36-Item Short Form Health Survey; SF-36 (Q3.8), Question 3.8 from 36-Item Short Form Health Survey.; SF-36 (Q3.9), Question 3.9 from 36-Item Short Form Health Survey.; SSQ Physical Function, Physical Function Scale of The Swiss Spinal Stenosis Questionnaire; Swiss.PF_Walk Item, Walking Distance Item from The Physical Function Scale Of The Swiss Spinal Stenosis Questionnaire; Swiss.SS_Balance, Balance Item from Symptom Severity Scale Of The Swiss Spinal Stenosis Questionnaire; Swiss.SS_Weak, Weak Item from Symptom Severity Scale Of The Swiss Spinal Stenosis Questionnaire

3.7.6 Table 8: Summary of results for known groups validity

Performance Based Measurements PBMs	Known-Group	Studies	Back Pain Type	Results Ratings *p*-value ≤ 0.05	COSMIN ROB	Result Rating per PBM	Level of Evidence
Stair Climbing (Time)	LBP - No LBP	Pfingsten, M., et al. (2014)	Non-Specific LBP	+	Very Good	+	Strong
Roll-up test	LBP - No LBP	Pfingsten, M., et al. (2014)	Non-Specific LBP	+	Very Good	+	Strong
Stand to Floor Tests	LBP - No LBP	Pfingsten, M., et al. (2014)	Non-Specific LBP	+	Very Good	+	Strong
Sock Tests	LBP - No LBP	Pfingsten, M., et al. (2014)	Non-Specific LBP	+	Very Good	+	Strong
Pick-up test	LBP - No LBP	Pfingsten, M., et al. (2014)	Non-Specific LBP	+	Very Good	+	Strong
5-Repetition Lift Test (ordinal 0-3)	LBP - No LBP	Pfingsten, M., et al. (2014)	Non-Specific LBP	+	Very Good	+	Strong
Back Performance Scale	Pain Level: -High Pain NRS ≥ 4 -Low Pain < 4)	Engh, L., et al. (2015)	Mixed LBP population	+	Doubtful	+	Limited
	Self-Reported Activity Level: -High Activity: not reduced, slightly reduced -Low Activity: fairly, very reduced	Engh, L., et al. (2015)	Mixed LBP population	+	Doubtful	+	Limited
	Work Ability: employed vs sick leave	Engh, L., et al. (2015)	Mixed LBP population	–	Doubtful	–	Limited

3.7.7 Table 9: Summary of results for Responsiveness - Construct Approach (hypothesis testing: comparison with other PROMs)

Performance-Based Measurements	PROMs	Studies	Back Pain Type	Score	Results Ratings	COSMIN ROB	PBM Overall Result Rating	Level of Evidence
5-Minute Walk Test (min)	COPM walk (p) Baseline to post-treatment	Andrew Walsh, D., et al. (2004)	Mixed LBP population	0.35	–	Very Good	–	Limited
	COPM walk (p) Baseline to 9-months follow up	Andrew Walsh, D., et al. (2004)	Mixed LBP population	0.24	–	Very Good	–	Limited
	COPM walk (s) Baseline to post-treatment	Andrew Walsh, D., et al. (2004)	Mixed LBP population	0.24	–	Very Good	–	Limited
	COPM walk (s) Baseline to 9-months follow up	Andrew Walsh, D., et al. (2004)	Mixed LBP population	0.18	–	Very Good	–	Limited
	ODI	Jakobsson, M., et al. (2019)	Degenerative LBP	–0.422	–	Very Good	–	Limited
50-ft WT (s) Time	ODI	Jakobsson, M., et al. (2019)	Degenerative LBP	0.467	–	Very Good	–	Limited
50-ft WT (m/s) Speed	FFbH-R	Strand, L. I., et al. (2011)	Non-Specific LBP	0.26	–	Very Good	–	Limited
	RMDQ	Strand, L. I., et al. (2011)	Non-Specific LBP	0.14	–	Very Good	–	Limited
TUG	ODI	Jakobsson, M., et al. (2019)	Degenerative LBP	0.413	–	Very Good	–	Limited

Self-Paced Walk Test (Time)	Estimated walking time	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.07	–	Doubtful	–	Limited
	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.08	–	Doubtful	–	Limited
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.17	–	Doubtful	–	Limited
	SSQ Physical Function Scale	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.25	–	Doubtful	–	Limited
Self-Paced Walk Test (Distance)	ODI	Tomkins-Lane, C. C., et al. (2014)	Lumber Spinal Stenosis	–0.70 (–0.93-0.25)	+	Very Good	+	Strong
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.23	–	Doubtful	–	Limited
		Tomkins-Lane, C. C., et al. (2014)	Lumber Spinal Stenosis	–0.78 (–1.04-0.50)	+	Very Good		
	SSQ Physical Function Scale	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.36	–	Doubtful	–	Limited
		Tomkins-Lane, C. C., et al. (2014)	Lumber Spinal Stenosis	–0.56 (–0.90-0.19)	+	Very Good		
	SSQ-PF Item 1 (distance)	Tomkins-Lane, C. C., et al. (2014)	Lumber Spinal Stenosis	–0.50 (–0.90-0.20)	+	Very Good	+	Strong
	Self-reported walking capacity	Tomkins-Lane, C. C., et al. (2014)	Lumber Spinal Stenosis	0.78 (0.46-1.02)	+	Very Good	+	Strong
	Estimated walking time	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.09	–	Doubtful	–	Limited

	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.11	–	Doubtful	–	Limited
MTT-Time	Estimated walking time	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.48	–	Doubtful	–	Limited
	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.37	–	Doubtful	–	Limited
	SSQ Physical Function Scale	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.35	–	Doubtful	–	Limited
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.48	–	Doubtful	–	Limited
MTT-Distance	Estimated walking time	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.50	+	Doubtful	+	Limited
	Estimated walking distance	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.62	+	Doubtful	+	Limited
	SSQ Physical Function Scale	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.41	–	Doubtful	–	Limited
	ODI-walking	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.35	–	Doubtful	–	Limited
Sock test	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.28	–	Doubtful	–	Limited
Sit-up test	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.53	+	Doubtful	+	Limited
Stand to floor	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.26	–	Doubtful	–	Limited

5-Repetition Lift Test (ordinal 0-4)	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.27	–	Doubtful	–	Limited
20 Steps–Stair Climbing (Time)	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.23	–	Doubtful	–	Limited
Pick-up test	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.04	–	Doubtful	–	Limited
Functional test index score	RMDQ	Caporaso, F., et al. (2012)	Non-Specific LBP	0.55	+	Doubtful	+	Limited
PWPE-Overall score	ODI	Durand, M. J., et al. (2008)	Non-Specific LBP	–0.28	–	Doubtful	–	Limited
	FABQ-PA	Durand, M. J., et al. (2008)	Non-Specific LBP	–0.16	–	Doubtful	–	Limited
	FABQ-work	Durand, M. J., et al. (2008)	Non-Specific LBP	–0.16	–	Doubtful	–	Limited
	PDI	Durand, M. J., et al. (2008)	Non-Specific LBP	0.07	–	Doubtful	–	Limited
PILE (highest load, kg)	FFbH-R	Strand, L. I., et al. (2011)	Non-Specific LBP	0.22	–	Very Good	–	Limited
	RMDQ	Strand, L. I., et al. (2011)	Non-Specific LBP	0.20	–	Very Good	–	Limited
Lift test (no. of lifts in 1 min)	FFbH-R	Strand, L. I., et al. (2011)	Non-Specific LBP	0.31	–	Very Good	–	Limited
	RMDQ	Strand, L. I., et al. (2011)	Non-Specific LBP	0.18	–	Very Good	–	Limited
BPS	FFbH-R	Strand, L. I., et al. (2011)	Non-Specific LBP	0.43	–	Very Good	–	Limited

	RMDQ	Strand, L. I., et al. (2011)	Non-Specific LBP	0.25	–	Very Good	–	Limited

ODI, Oswestry Disability Index; ODI-walking, Walking Distance Item from The Oswestry Disability Index; RMDQ, Roland-Morris Disability Questionnaire; COPM walk (p), Reported Performance with Walk Tolerance from Canadian Occupational Performance Measure; COPM walk (s), Reported Satisfaction with Walk Tolerance from Canadian Occupational Performance Measure; FFbH-R, Hannover Functional Ability Questionnaire for Measuring Back Pain-Related Disability; SSQ Physical Function, Physical Function Scale of The Swiss Spinal Stenosis Questionnaire; FABQ-PA and FABQ-work, Physical Activity and Work Subscales from Fear-avoidance Beliefs Questionnaire

3.7.8 Table 10: Summary of results for Responsiveness - Construct Approach (hypothesis testing: comparison between subgroups)

Discrimination Between Groups	Performance Based Measurements PBM	Studies	Back Pain Type	Score	Results Ratings *p*-value ≤ 0.05	COSMIN ROB	Result Rating per PBM	Level of Evidence
(very much, much, or slightly improved versus no change) *Linear Trend, Contrast and P Values*	Progressive Isoinertial Lifting Evaluation	Strand, L. I., et al. (2011)	Non-Specific LBP	*–5.85* *p*-value: 0.076	–	Very Good	–	Limited
	Lift Test	Strand, L. I., et al. (2011)	Non-Specific LBP	*–0.48, P=0.006*	+	Very Good	+	Strong
	15-m walk test (m/s)	Strand, L. I., et al. (2011)	Non-Specific LBP	0.24, P=0.212	–	Very Good	–	Limited
	Back Performance Scale	Strand, L. I., et al. (2011)	Non-Specific LBP	–6.35, P< –0.001	+	Very Good	+	Strong
better score, worse score *15-point scale ranging from -7 to +7*	Physical Work Performance Evaluation (Overall score)	Durand, M. J., et al. (2008)	Non-Specific LBP	7+; 1- *p*-value: 0.0606	–	Doubtful	–	Limited

3.7.9 Table 11: Summary of results for Responsiveness – Criterion Approach (comparison to a gold standard) using Generic-Global Rating Scale (Unchanged-Improved)

Performance Based Measurements PBMs vs. Generic-GRS (Unchanged-Improved)	Studies	Back Pain Type	AUC (95% CI)	Results Rating AUC ≥ 0.70	COSMIN ROB	Overall Result Rating per PBM	Level Of Evidence
5-Repetition Sit to Stand (s)	Andersson, E. I., et al. (2010)	Non-Specific LBP	0.75 (0.66 to 0.83)	+	Inadequate	+	Poor
Stair Climbing (steps)	Andersson, E. I., et al. (2010)	Non-Specific LBP	0.72 (0.62 to 0.81)	+	Inadequate	+	Poor
	Jakobsson, M., et al. (2019)	Degenerative LBP	0.70 (0.59 to 0.81)	+	Very Good	+	Strong
5-Minute Walk Test (m)	Andersson, E. I., et al. (2010)	Non-Specific LBP	0.60 (0.50 to 0.69)	–	Inadequate	–	Poor
	Jakobsson, M., et al. (2019)	Degenerative LBP	0.70 (0.58 to 0.82)	+	Very Good	+	Strong
50-ft Walk Test (s)	Andersson, E. I., et al. (2010)	Non-Specific LBP	0.64 (0.54 to 0.74)	–	Inadequate	–	Poor
	Jakobsson, M., et al. (2019)	Degenerative LBP	0.76 (0.66 to 0.87)	+	Very Good	+	Strong
Time Up and Go (s)	Jakobsson, M., et al. (2019)	Degenerative LBP	0.72 (0.67 to 0.91)	+	Very Good	+	Strong
	Jakobsson, M., et al. (2019)	Degenerative LBP	0.72 (0.62 to 0.83)	+	Very Good	+	Strong
Shuttle Walking Test (distance)	Campbell, H., et al. (2006).	Mixed LBP population	0.77 (0.71 to 0.83)	+	Very Good	+	Strong

Self-Paced Walking Test DISTANCE	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.564	–	Doubtful	–	Limited
Self-Paced Walking Test TIME	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.545	–	Doubtful	–	Limited
Motorized Treadmill Test DISTANCE	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.702	+	Doubtful	+	Limited
Motorized Treadmill Test TIME	Rainville, J., et al. (2012)	Lumber Spinal Stenosis	0.717	+	Doubtful	+	Limited
Progressive Isoinertial Lifting Evaluation Test (cycles)	Andersson, E. I., et al. (2010)	Non-Specific LBP	0.59 (0.49 to 0.69)	–	Inadequate	–	Poor
Lifting Test (no. of lifts in 1 min)	Strand, L. I., et al. (2011)	Non-Specific LBP	0.64 (0.52 to 0.76)	–	Doubtful	–	Limited
Back Performance Scale (scale 0–15)	Strand, L. I., et al. (2011)	Non-Specific LBP	0.73 (0.60 to 0.86)	+	Doubtful	+	Limited

3.7.10 Table 12: Summary of results for Responsiveness - Criterion Approach (comparison to a gold standard) using Specific-Global Rating Scale (Unchanged-Improved)

Performance Based Measurements PBMs	Responsiveness - Criterion Approach Specific-GRS	Article	Back Pain Type	AUC (95% CI)	Results Rating AUC ≥ 0.70	COSMIN ROB	Overall Result Rating per PBM	Level Of Evidence
Stair Climbing (steps)	GRS-Stair Climbing Unchanged-Improved	Jakobsson, M., et al. (2019)	Degenerative LBP	0.72 (0.59 to 0.85)	+	Very Good	+	Strong
5-Minute Walk Test (m)	GRS-walking Unchanged-Improved	Jakobsson, M., et al. (2019)	Degenerative LBP	0.68 (0.54 to 0.82)	–	Very Good	–	Limited
50-ft Walk Test (s)	GRS-walking Unchanged-Improved	Jakobsson, M., et al. (2019)	Degenerative LBP	0.80 (0.67 to 0.93)	+	Very Good	+	Strong
Time Up and Go (s)	GRS-walking Unchanged-Improved	Jakobsson, M., et al. (2019)	Degenerative LBP	0.74 (0.61 to 0.86)	+	Very Good	+	Strong
Time Up and Go (s)	GRS-chair rise Unchanged-Improved	Jakobsson, M., et al. (2019)	Degenerative LBP	0.79 (0.67 to 0.91)	+	Very Good	+	Strong
Self-Paced Walk Test DISTANCE	GRS-Walking Unchanged-Improved	Tomkins-Lane, C. C., et al. (2014)	Lumber Spinal Stenosis	0.92	+	Doubtful	+	Limited

3.7.11 Table 13: Summary of results for Psychometric Properties of PBMs.

Performance-Based Measurement	Back Pain Classifications	Psychometric Properties							
		Reliability			Validity		Responsiveness		
		Test-Retest	Inter-Rater	Intra-Rater	Convergent	Known-Groups	Criterion Approaches General-GPE	Criterion Approaches Specific-GPE	Construct Approaches
Walking Test (8 walk tests)									
5-Minute Walk Test (m)	Non-specific Low Back Pain	Limited (+) by 1 study	-	-	correlated-RMDQ Limited (–) by 1 study	-	Poor (–) by 1 study	-	-
	Mixed Low Back Pain population	Limited (+) by 1 study	-	-	correlated-RMDQ Limited (–) by 3 studies correlated-COPM walk (performance) Limited (–) by 1 study correlated-COPM walk (satisfaction) Limited (–) by 1 study correlated-Self-Reported PA Limited (–) by 1 study	-	-	-	correlated-COPM walk (P+S) T0, T1, T2 Limited (–) by 1 study
	Low Back Pain due to degenerative changes	-	-	-	-	-	Strong (+) by 1 study	Limited (–) by 1 study	correlated-ODI Limited (–) by 1 study
6-Minute Walk Test (m)	Lumber Spinal Stenosis	-	-	-	Correlated-ODI Limited (–) by 1 study Correlated-Quebec Limited (–) by 1 study	-	-	-	-
50-Feet Walk Test (s)	Non-specific Low Back Pain	Moderate (+) by 3 studies	Limited (+) by 1 study	Limited (+) by 1 study	correlated-RMDQ Limited (–) by 2 studies	-	Poor (–) by 1 study	-	-
	Mixed Low Back Pain population	Limited (+) by 1 study	-	-	correlated-RMDQ Limited (–) by 3 studies	-	-	-	-

	Low Back Pain due to degenerative changes	-	-	-	-	-	Strong (+) by 1 study	Strong (+) by 1 study	correlated-ODI Limited (–) by 1 study
Shuttle Walk Test (m)	Non-specific Low Back Pain	Limited (+) by 1 study	-	-	-	-	Strong (+) by 1 study	-	-
Timed Up and Go Test (s)	Non-specific Low Back Pain	Limited (+) by 1 study	Limited (+) by 1 study	Limited (+) by 1 study	-	-	-	-	-
	Lumber Spinal Stenosis	-	-	-	Correlated-ODI Limited (–) by 1 study Correlated-PFS Strong (+) by 1 study	-	-	-	-
	Low Back Pain due to degenerative changes	-	-	-	-	-	Strong (+) by 1 study	Strong (+) by 1 study	correlated-ODI Limited (–) by 1 study
	Mixed Low Back Pain population	Limited (+) by 1 study	-	-	Correlated-RMDQ Limited (–) by 1 study	-	-	-	-
Self-Paced Walking Test (Distance)	Lumber Spinal Stenosis	Limited (+) by 1 study	-	-	Correlated-SSQ (physical Function) Strong (+) by 3 study Correlated-Quebec Strong (+) by 1 study Correlated-SF-36 (physical function) Strong (+) by 1 study Correlated-ODI Strong (+) by 2 study	-	Limited (–) by 1 study	Limited (+) by 1 study	correlated-ODI Strong (+) by 1 study correlated-Self- Walk Cap Strong (+) by 1 study

					Correlated-Self-Estimated Walk "Distance" Strong (+) by 3 study Correlated-Self-Estimated Walk Cap "ordinal 0-10" Strong (+) by 2 study Correlated-Self-Estimated Walk "Time" Moderate (+) by 1 study				correlated-SSQ (physical Function) Limited (–) by 2 study correlated-Self- Walk Estimates "Distance & Time" Limited (–) by 1 study
Self-Paced Walking Test (TIME)	Lumber Spinal Stenosis	-	-	-	Correlated-SSQ (physical Function) Moderate (+) by 1 study Correlated-Self-Estimated Walk "Distance" Moderate (+) by 1 study Correlated-Self-Estimated Walk "Time" Moderate (+) by 1 study	-	Limited (–) by 1 study	-	correlated-SSQ (physical Function) Limited (–) by 2 study correlated-Self- Walk Estimates "Distance & Time" Limited (–) by 1 study
15-meter Walk Test (m/s)	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study	-	-	-	Correlated-RMDQ Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study
Free Walking Velocity Test (speed)	Lumber Spinal Stenosis	-	-	-	Correlated-ODI Moderate (+) by 1 study Correlated-Quebec Moderate (+) by 1 study	-	-	-	- -
Treadmill Walking Tests (6 tests)									
Deen-Treadmill Walk Test (total Ambulation time) 1.2 speed	Lumber Spinal Stenosis	Poor (+) by 1 study	-	-	-	-	-	-	-

Deen- Treadmill Walk Test (total Ambulation time) preferred speed	Lumber Spinal Stenosis	Poor (+) by 1 study	-	-	-	-	-	-	-
Ambulatory- Treadmill test (distance)	Lumber Spinal Stenosis	-	-	-	Correlated- ODI Strong (+) by 1 study Correlated- Patient Walk Expectation (Distance) Strong (+) by 1 study	-	-	-	-
Modified- Symptom- Limited Treadmill Test (time)	Mixed Low Back Pain population	-	-	-	correlated- SF-36 (physical functioning domain) Limited (–) by 1 study	-	-	-	-
Motorised Treadmill Test (Time)	Lumber Spinal Stenosis	-	-	-	correlated- SSQ Physical Function Moderate (+) by 1 study correlated- Self- Estimated Walk "Distance" Moderate (+) by 1 study correlated- Self- Estimated Walk "Time" Moderate (+) by 1 study	-	Limited (+) by 1 study	-	correlated- SSQ (physical Function) Limited (–) by 2 study correlated- Self- Walk Estimates "Distance & Time" Limited (–) by 1 study
Motorised Treadmill Test (Distance)	Lumber Spinal Stenosis	-	-	-	correlated- SSQ Physical Function Limited (–) by 1 study correlated- Self- Estimated Walk "Distance" Moderate (+) by 1 study correlated- Self- Estimated Walk "Time" Moderate (+) by 1 study	-	Limited (+) by 1 study	-	correlated- SSQ (physical Function) Limited (–) by 2 study correlated- Self- Walk Estimates "Distance & Time" Limited (+) by 1 study

Treadmill Tolerance Test (Time)	Lumber Spinal Stenosis	-	-	-	correlated-ODI Strong (+) by 1 study correlated-SSQ Physical Function Strong (+) by 1 study	-	-	-	-
Felix-Treadmill Walk Test	Lumber Spinal Stenosis	-	-	-	Correlated-Self-Estimated Walk "Distance" Limited (–) by 1 study	-	-	-	-
Sit-to-Sand Tests									
5-Repetition Sit-to-Stand Test (time)	Non-specific Low Back Pain	Limited (–) by 2 study	Limited (+) by 1 study	Limited (+) by 1 study	Correlated-RMDQ Limited (–) by 2 study	-	Poor (+) by 1 study	-	-
	Mixed Low Back Pain population	Limited (+) by 1 study	-	-	Correlated-RMDQ Limited (–) by 3 study Correlated-Self-Reported PA Limited (–) by 1 study	-	-	-	-
	Low Back Pain due to degenerative changes	Poor (+) by 1 study	-	-	Correlated-RMDQ Limited (–) by 1 study Correlated-ODI Limited (–) by 1 study Correlated-EQ-5D index Strong (+) by 1 study	-	-	-	-
30s-Chair Stand Test	Non-specific Low Back Pain	-	-	Limited (+) by 1 study	Correlated-ODI Limited (–) by 1 study	-	-	-	-
Stair Climbing Tests									

1-Minute Stair Climbing (steps)	Non-specific Low Back Pain	Limited (+) by 1 study	-	-	-	-	Poor (+) by 1 study	-	-
	Low Back Pain due to degenerative changes	-	-	-	-	-	Strong (+) by 1 study	Strong (+) by 1 study	-
	Mixed Low Back Pain population	-	Limited (–) by 1 study	Limited (–) by 1 study	-	-	-	-	-
20 Steps–Stair Climbing (Time)	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study	Subgroups (LBP-No LBP) Strong (+)	-	-	Correlated-RMDQ Limited (–) by 1 study
Lateral Step Down	Non-specific Low Back Pain	-	-	Limited (+) by 1 study	-	-	-	-	-
Lifting Tests (12 tests)									
BPS-Lifting Test	Mixed Low Back Pain population	Limited (–) by 2 study	Limited (–) by 1 study	Limited (+) by 1 study	-	-	-	-	-
Lift test (no. of lifts in 1 min)	Non-specific Low Back Pain	Poor (+) by 1 study	-	-	Correlated-FFbH-R Limited (–) by 1 study Correlated-RMDQ Limited (–) by 1 study	-	Limited (–) by 1 study	-	Correlated-RMDQ Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study
FCE-safe maximum lifting	Mix: Specific and Non-specific Low Back Pain	-	Limited (–) by 1 study	Moderate (+) by 1 study	-	-	-	-	-
IWS FCE-Lifting Test Borg CR-10 scale	Non-specific Low Back Pain	-	Strong (+) by 1 study	-	-	-	-	-	-

IWS FCE-Lifting Test Categorical Scale	Non-specific Low Back Pain	-	Limited (–) by 1 study	-	-	-	-	-	-
Back-torso lift test	Lumber Spinal Stenosis	-	Strong (+) by 1 study	-	-	-	-	-	-
Shoulder lift test	Lumber Spinal Stenosis	-	Strong (+) by 1 study	-	-	-	-	-	-
Carrying lifting strength test	Lumber Spinal Stenosis	-	Strong (+) by 1 study	-	-	-	-	-	-
Lower lifting strength test	Lumber Spinal Stenosis	-	Strong (+) by 1 study	-	-	-	-	-	-
Upper lifting strength test	Lumber Spinal Stenosis	-	Strong (+) by 1 study	-	-	-	-	-	-
5-Repetition Lift Test (ordinal 0-4)	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study	Subgroups (LBP-No LBP) Strong (+)	-	-	Correlated-RMDQ Limited (–) by 1 study
ADL Tests (10 tests)									
BPS-Sock Test	Mixed Low Back Pain population	Limited (–) by 2 study	Limited (–) by 1 study	Limited (+) by 1 study	-	-	-	-	-
	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study	Subgroups (LBP-No LBP) Strong (+)	-	-	Correlated-RMDQ Limited (–) by 1 study
BPS-Pick Up Test	Mixed Low Back Pain population	Limited (–) by 2 study	Limited (–) by 1 study	Limited (–) by 1 study	-	-	-	-	-

	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study	Subgroups (LBP-No LBP) Strong (+)	-	-	Correlated-RMDQ Limited (–) by 1 study
BPS-Rolling Test	Mixed Low Back Pain population	Limited (–) by 2 study	-	-	Correlated-RMDQ Limited (–) by 1 study Correlated-Self-Reported PA Limited (–) by 1 study	-	-	-	-
360 Roll-Over Test	Mixed Low Back Pain population	-	-	-	Correlated-RMDQ Limited (–) by 1 study	-	-	-	-
Roll-Up Test	Non-specific Low Back Pain	-	-	-	-	Subgroups (LBP-No LBP) Strong (+)	-	-	-
Rising–Up Test	Mixed Low Back Pain population	-	Limited (–) by 1 study	Limited (–) by 1 study	-	-	-	-	-
Lacing Test	Mixed Low Back Pain population	-	Limited (–) by 1 study	Limited (+) by 1 study	-	-	-	-	-
Sit-Up test	Mixed Low Back Pain population	-	Limited (–) by 1 study	Limited (–) by 1 study	-	-	-	-	-
	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study	-	-	-	Correlated-RMDQ Limited (+) by 1 study

Hair-Wash Test	Mixed Low Back Pain population	-	Limited (–) by 1 study	Limited (+) by 1 study	-	-	-	-	-
Stand-to-Floor Test	Mixed Low Back Pain population	-	Limited (–) by 1 study	Limited (–) by 1 study	-	-	-	-	-
	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Limited (–) by 1 study	Subgroups (LBP-No LBP) Strong (+)	-	-	Correlated-RMDQ Limited (–) by 1 study
Balance Tests									
One Leg Stand Test	Mixed Low Back Pain population	-	Limited (–) by 2 study	Moderate (+) by 3 study	Correlated-ODI Limited (–) by 1 study Correlated-PFS Limited (–) by 1 study	-	-	-	-
Battery Tests (7 tests)									
Back Performance Scale	Non-specific Low Back Pain	Poor (+) by 1 study	-	-	Correlated-FFbH-R Strong (+) by 1 study Correlated-RMDQ Limited (–) by 1 study	-	Limited (+) by 1 study	-	Correlated-RMDQ Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study
	Mixed Low Back Pain population	Moderate (+) by 2 studies	-	-	Correlated-ODI Limited (+) by 1 study Correlated-RMDQ Limited (+) by 1 study	Subgroups (High Pain - Low Pain Subgroups (High Activity - Low Activity) Subgroups (employed - sick leave) Limited (+)	-	-	-

	Lumber Spinal Stenosis	-	-	-	Correlated-FFbH-R Strong (+) by 1 study	-	-	-	-
Function Battery Test	Mixed Low Back Pain population	-	Limited (+) by 1 study	Limited (+) by 1 study	-	-	-	-	-
Functional Test Index Score	Non-specific Low Back Pain	-	-	-	Correlated-RMDQ Strong (+) by 1 study	-	-	-	Correlated-RMDQ Limited (+) by 1 study
Progressive Isoinertial Lifting Evaluation (highest load, kg)	Muscular-related Low Back Pain	Limited (+) by 1 study	-	-	-	-	-	-	-
	Non-specific Low Back Pain	Limited (+) by 2 studies	-	-	Correlated-RMDQ Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study	-	Poor (–) by 1 study	-	Correlated-RMDQ Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study
Villiger Test (No. of steps)	Mixed Low Back Pain population	-	-	-	Correlated-PDI Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study	-	-	-	-
Villiger Test (Test Duration)	Mixed Low Back Pain population	-	-	-	Correlated-PDI Limited (–) by 1 study Correlated-FFbH-R Limited (–) by 1 study	-	-	-	-
Functional Capacity Evaluation	Non-specific Low Back Pain	-	-	-	Correlated-ODI Strong (+) by 1 study Correlated-Quebec Strong (+) by 1 study Correlated-RMDQ Limited (–) by 1 study	-	-	-	-

Physical Work Performance Evaluation	Non-specific Low Back Pain	-	-	-	-	-	-	-	correlated-ODI Limited (–) by 1 study	correlated-PDI Limited (–) by 1 study

COPM walk (performance), Reported Performance With Walk Tolerance from Canadian Occupational Performance Measure; COPM walk (satisfaction), Reported Satisfaction With Walk Tolerance from Canadian Occupational Performance Measure; EQ-5D (Q 1), Mobility Item from Euro-Quality Of Life 5 Domains; EQ-5D index, Euro-Quality Of Life 5 Domains; FFbH-R, Hannover Functional Ability Questionnaire for Measuring Back Pain-Related Disability; HUI3 Amb, Health Utilities Index Single Attribute Utility Score for Ambulation; ODI, Oswestry Disability Index; ODI-walking, Walking Distance Item from The Oswestry Disability Index; PDI, Pain Disability Index; PFS, Physical Functional Scale; Quebec, Quebec Back Pain Disability Scale; Quebec_Groceries, Groceries Item from Quebec Back Pain Disability Scale; Quebec_Reach, Reach Item from Quebec Back Pain Disability Scale; Quebec_Run, Run Item from Quebec Back Pain Disability Scale; Quebec_Stand, Stand Item from Quebec Back Pain Disability Scale; Quebec_Walk, Walk Item from Quebec Back Pain Disability Scale; Quebec-8, Question 8 from Quebec Back Pain Disability Scale; Quebec-9, Question 9 from Quebec Back Pain Disability Scale; RMDQ, Roland-Morris Disability Questionnaire; RMDQ-17, Question 17 from Roland-Morris Disability Questionnaire; RMDQ-3, Question 3 from Roland-Morris Disability Questionnaire; SF-36 (physical functioning domain), Physical Functioning Domain from 36-Item Short Form Health Survey; SF-36 (Q3.7), Question 3.7 from 36-Item Short Form Health Survey; SF-36 (Q3.8), Question 3.8 from 36-Item Short Form Health Survey.; SF-36 (Q3.9), Question 3.9 from 36-Item Short Form Health Survey.; SSQ Physical Function, Physical Function Scale of The Swiss Spinal Stenosis Questionnaire; Swiss.PF_Walk Item, Walking Distance Item from The Physical Function Scale Of The Swiss Spinal Stenosis Questionnaire; Swiss.SS_Balance, Balance Item from Symptom Severity Scale Of The Swiss Spinal Stenosis Questionnaire; Swiss.SS_Weak, Weak Item from Symptom Severity Scale Of The Swiss Spinal Stenosis Questionnaire

Chapter Four: Discussion and Conclusion

4.1 Introduction

The current thesis focused on the psychometric properties of performance-based measures (PBMs) used or developed to assess physical function in low back pain (LBP). Chapter one described LBP prevalence, risk factors, prognosis and management. It included a brief introduction about the core outcome measurement set recommended for assessing physical function in LBP studies, which focuses primarily on patient-reported outcome measures (PROMs). Psychometric properties of PROMs were highlighted, in addition to the evidence on the psychometric properties of PBMs used in LBP. Chapter two presented the systematic review protocol that was then presented in chapter three in the format for submission to the Journal of Orthopedic and Sports Physical Therapy. The review followed COnsensus-based Standards for the selection of health status Measurement INstrument (COSMIN) standards for conducting systematic reviews of outcome measures.[1] Risk of bias of included studies was assessed using the COSMIN Risk of Bias checklist 2018 (COSMIN-ROB).[2] Level of evidence of the identified PBMs were synthesized following GRADE standards mentioned in the COSMIN handbook (2018).[1] Chapter four included a lay summary, key findings, strengths and limitations, in addition to implications for clinical practice and future research.

4.2 Lay summary

The current study was carried out to identify tests that we can use to evaluate the body's ability to perform physical activities, such as self-care, walking, or stair climbing, in people with low back pain. We were also interested in how accurate and reliable these tests were. There were three features that we looked at to determine how useful these tests were. The first feature (reliability) was related to the consistency of the test every time we use it. The second feature (validity) was concerned with the soundness of what the test was evaluating. Validity allows us to evaluate if a test used to examine physical ability is actually examining physical ability and not some other outcome. The last feature (responsiveness) was concerned about the test's ability to capture changes over time in patients that had really changed.

This review found 47 studies (with 115 physical tests) that evaluated the features mentioned above. None of these tests had their three features reviewed completely. For example, some tests had good validity but not very good reliability or were not tested for responsiveness. This is why we could not advise on one ideal test. In addition, most of these 47 studies were not considered good quality, and because of this, their results may not be trustworthy. In conclusion, we need more high-quality studies that evaluate physical tests' characteristic before they can be recommended for use in clinical practice and research in people with LBP. The current advice is for clinicians and researchers to use these tests in combination with other well-tested questionnaires.

4.3 Key Findings

There were 47 studies that met the inclusion criteria,[3-49] with five LBP diagnoses (e.g., non-specific LBP, spinal stenosis) and different LBP duration (e.g., acute, chronic). In general, findings included the following:

1. Most of the levels of evidence were generated from single studies for each PBM or psychometric property.
2. The majority of the included studies had a high risk of bias assessed by the COSMIN-ROB checklist.[2]
3. A large number of studies did not find PBMs to have good psychometric properties as results/scores did not meet the pre-defined thresholds/hypothesis for good psychometrics.

4. The great majority of PBMs' psychometric properties were found to have a low level of evidence.

Specifically, for each psychometric property, key findings included:

A. *Reliability:*

   a. Test-retest reliability: the 50-Feet Walk Test and Back Performance Scale demonstrated "Moderate" level of evidence generated from two or more studies.

   b. Inter-rater reliability: the Back-Torso Lift Test, Shoulder Lift Test, Carrying Lifting Strength Test, Lower Lifting Strength Test, and Upper Lifting Strength Test demonstrated "Strong" level of evidence due to the high quality of their study, and the results met the pre-defined hypothesis.

   c. Intra-rater reliability: One Leg Stand Test and Functional Capacity Evaluation (safe maximum lifting) had a "Moderate" level of evidence. The remaining PBMs had a low level of evidence, mainly due to the included studies' low quality.

B. *Convergent Validity:*

   a. Although most validity studies were of high quality, the majority of PBMs had a low level of evidence due to not meeting the pre-defined hypotheses for good validity.

   b. PBMs of "Strong" level of evidence generated from single studies per each test: Shuttle Walk Test (Distance), Self-Paced Walking Test (Distance), Timed Up and Go Test, Ambulatory-Treadmill test (distance), Treadmill Tolerance Test (Time), Back Performance Scale, Function Battery Test, and Functional Capacity Evaluation.

   c. PBMs of "Strong" level of evidence generated from two or more studies: Self-Paced Walking Test (Distance).

C. *Responsiveness*

a. PBMs of "Strong" level of evidence: Self-Paced Walk Test (Distance), 5-Minute Walk Test, 50-Feet Walk Test, Shuttle Walk Test (m), Timed Up and Go test, and 1-Minute Stair Climbing (steps).

Most of these PBMs' levels of evidence were generated from low-risk single studies per PBM.

4.4 Strengths and Limitations

A strength of this study was the use of the updated version of the COSMIN systematic review methodology manual (2018).[1] The updated manual recommends the inclusion of low-quality studies into data synthesis, which prevents exclusion of eligible studies while still considering their methodological limitations within the review conclusions.[1] A significant strength of the review was the use of pre-defined hypotheses for each psychometric property (mainly for validity and responsiveness), which allowed for the standardization of comparisons across all studies.[1] Furthermore, the WHO-ICF model, with consideration of the overlap between *Activity* and *Participation* domains, was used as a framework to define physical function.[50] Other strengths included providing results specific to LBP diagnoses and including non-English studies.

A limitation of this study was that most evidence was generated from single studies. This indicates that there is a lack of research and evidence for PBMs in LBP. An additional limitation of this review was excluding studies (3 studies) that lacked descriptions of PBMs' protocols and lacked a measurement unit or total score (e.g., kg). Also, informal methods for translation of non-English studies (3 studies) were used (e.g., www.translate.google.com), which might have led to misinterpretations. Given the heterogeneity of the included studies in terms of tests, population, and psychometric properties evaluated, no clear PBM could be recommended for use in clinical practice and research; the ultimate choice of PBM will be dependent on the context and purpose of the assessment. For example, if we want to examine changes in patients' physical status (e.g., physical function) after a particular treatment, we have to have enough evidence on the measurement's responsiveness. In addition, external factors such as equipment and space required for the measurement to be carry out, clinician familiarity with the measurement, or time required to complete the examination need to be taken in consideration.

4.5 Impact of the present study

Clinical Applicability

There was not a single PBM that demonstrated a good level of evidence for all psychometric properties. Hence, clinicians should be cautious when selecting and interpreting PBMs' outcomes in clinical practice. Further, PBMs should not be used alone, rather in combination with reliable PROMs. When selecting a PBM, clinicians need to make sure that the measure has been tested on the patient population in which they plan to use it as psychometric properties may change based on population characteristics. According to our results, some PBMs demonstrated different levels of evidence when used with Lumbar Spinal Stenosis as opposed to other LBP diagnoses. For example, the 50-ft Walk Test demonstrated poor responsiveness when used in non-specific LBP and had high responsiveness when used in degenerative LBP (older adults). In addition, clinicians need to make sure that the measure has been tested for the psychometric properties that suits their aims. For example, when a clinician is interested in observing change, then they need to consider a measure that has good evidence for responsiveness.

In general, the Self-Paced Walking Test specifically measured by distance was one of the PBMs that presented the best evidence for psychometric properties only in Lumbar Spinal

Stenosis patients. However, there no studies on its psychometric properties in other LBP diagnoses (e.g., non-specific LBP).

Recommendation for Future Research

There was considerable heterogeneity in the included studies in terms of tests, psychometric properties and population, which led to a limited level of evidence per PBM. Therefore, future research needs to focus on building upon existing evidence and filling the large gap of evidence concerning missing PBMs psychometric properties, especially in reliability and responsiveness. For example, the Self-Paced Walk Test demonstrated very good construct validity and responsiveness in assessing physical function in Lumbar Spinal Stenosis; however, there were no studies on its inter- and intra-reliability. Therefore, more studies on these properties with a focus on reliability are warranted. In addition, some of the PBMs had low levels of evidence because the studies were of low quality. Hence, we need high-quality studies to build onto this already available but limited evidence.[1]

There were no eligible studies on the criterion validity of the identified PBMs. This was expected as we do not have a gold standard to assess physical function in LBP. It was recently suggested that direct measurements such as Energy Expenditure Level or Actigraphy (e.g., accelerometer) could be used as a gold standard.[51] In general, direct measurements reflect the persons' metabolic cost or energy expenditure due to any physical activity that elevates heart rate beyond the resting levels.[52 53] However, these measurements often fail to detect the actual physical functioning, especially in the elderly.[52 53] Direct measurements are good at measuring activities that entail mobility and movement (changes in locomotion) but do not capture typical day to day activity (e.g., self-care, lifting); hence, the free-living functionality of a person.[52 53] This indicates a need for caution when interpreting direct measurements' outcomes used to assess physical function,[52 53] and a need for the development of advanced technology for detecting a person's physical function in all aspects of activity during daily living.

An interesting finding was that reliability studies were of poor quality due to the inclusion of participants undergoing interventions but were added into a reliability study if they reported "no change". Therefore, future studies that are designed explicitly to evaluate reliability are advised to include participants receiving no treatment.[54] Further, many reliability studies either chose the wrong model of Interclass Correlation coefficient (ICC) or did not report it. ICC models should be reported and should be specific to the type and purpose of reliability measurement.[54 55] Furthermore, SEM (Standard Error of Measurement) was often not reported. Therefore, future studies on reliability should use and present appropriate statistics, including absolute and relative reliability measures. [54 55]

Responsiveness was the least evaluated property in the included studies. An issue often observed was the use of general anchors (Global Rating Scale, GRS) that were not specific to physical function. Responsiveness studies should ideally include specific GRS related to the outcome being evaluated, such as questions on perceived change in physical function. [54] Further, prior to constructing ROC curves, it is important that the level of agreement between both measurements (GRS and PBM) is examined.[54] Among all the six studies that assessed responsiveness using the criterion approach, only two reported this information. Therefore, future studies on responsiveness should consider the inclusion of a construct specific GRS on physical function, as well as a priori evaluation of the agreement between GRS and PBM. [54]

4.6 Knowledge Translation

This thesis's results are expected to contribute to the selection of PBMs in both clinical practice and research. Researchers and clinicians will be able to use the comprehensive results of this thesis in selecting outcome measures and guiding future research in LBP. Furthermore, the systematic review will be published in a peer-reviewed journal, and results will be presented at national and international conferences focused on spine and rehabilitation.

4.7 Conclusion

There is limited evidence on the measurement properties of PBMs used to assess physical function in LBP patients; therefore, caution is recommended when using these measures in clinical practice and research. Moreover, there is a need for more high-quality studies that investigate the psychometric properties of PBMs of physical function in LBP. Promising PBMs were identified but need to be investigated in future studies such as, Self-Paced Walk Test (Distance) for Lumbar Spinal Stenosis; and 50-ft Walk Test (s), Timed Up and Go Test (s) and Back Performance Scale for all other diagnosis.

4.8 Reference

1. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018;78:1.
2. Mokkink L. COSMIN risk of bias checklist, 2018.
3. Alamam DM, Leaver A, Moloney N, et al. Pain Behaviour Scale (PaBS): An exploratory study of reliability and construct validity in a chronic low back pain population. *Pain Research and Management* 2019;2019 (no pagination)(2508019)
4. Andersson EI, Lin CC, Smeets R. Performance Tests in People With Chronic Low Back Pain Responsiveness and Minimal Clinically Important Change. *Spine* 2010;35(26):E1559-E63. doi: 10.1097/BRS.0b013e3181cea12e
5. Andrew Walsh D, Jane Kelly S, Sebastian Johnson P, et al. Performance problems of patients with chronic low-back pain and the measurement of patient-centered outcome. *Spine* 2004;29(1):87-93.
6. Armstrong M, McDonough S, Baxter D. Reliability and repeatability of shuttle walk test in patients with chronic low back pain...including commentary by Eiser N, Lemmink KAP, and Walsh DA. *International Journal of Therapy & Rehabilitation* 2005;12(10):438-43.
7. Barz T, Melloh M, Staub L, et al. The diagnostic value of a treadmill test in predicting lumbar spinal stenosis. *European Spine Journal* 2008;17(5):686-90. doi: 10.1007/s00586-008-0593-1
8. Campbell H, Rivero-Arias O, Johnston K, et al. responsiveness of objective, disease-specific, and generic outcome measures in patients with chronic low back pain: an assessment for improving, stable, and deteriorating patients. *Spine* 2006;31(7):815-22.

9. Caporaso F, Pulkovski N, Sprott H, et al. How well do observed functional limitations explain the variance in Roland Morris scores in patients with chronic non-specific low back pain undergoing physiotherapy? *European Spine Journal* 2012;21:S187-S95. doi: 10.1007/s00586-012-2255-6

10. Conway J, Tomkins CC, Haig AJ. Walking assessment in people with lumbar spinal stenosis: capacity, performance, and self-report measures. *Spine Journal* 2011;11(9):816-23. doi: 10.1016/j.spinee.2010.10.019

11. Cunha IT, Simmonds MJ, Protas EJ, et al. Back pain, physical function, and estimates of aerobic capacity - What are the relationships among methods and measures? *American Journal of Physical Medicine & Rehabilitation* 2002;81(12):913-20. doi: 10.1097/01.Phm.0000030729.77020.2a

12. da Cunha-Filho IT, Lima FC, Guimaraes FR, et al. use of physical performance tests in a group of Brazilian Portuguese-speaking individuals with low back pain. *Physiotherapy Theory & Practice* 2010;26(1):49-55. doi: 10.3109/09593980802602844

13. Deen HG, Zimmerman RS, Lyons MK, et al. Test-Retest Reproducibility of the Exercise Treadmill Examination in Lumbar Spinal Stenosis. *Mayo Clinic Proceedings* 2000;75(10):1002-07.

14. Denteneer L, van Daele U, Truijen S, et al. Convergent validity of clinical tests which are hypothesized to be associated with physical functioning in patients with non-specific chronic low back pain. *Journal of Back & Musculoskeletal Rehabilitation* 2019;16:16.

15. Durand MJ, Brassard B, Hong QN, et al. Responsiveness of the Physical Work Performance Evaluation, a functional capacity evaluation, in patients with low back pain. *Journal of Occupational Rehabilitation* 2008;18(1):58-67.

16. Engh L, Strand L, Robinson H, et al. Back Performance Scale: Assessment of patients with back problems in primary health care. *Fysioterapeuten* 2015;82(9):22-7.

17. Felix ZF, Schiltenwolf M, Abel R, et al. Gait analysis does not correlate with clinical and MR imaging parameters in patients with symptomatic lumbar spinal stenosis. *Bmc Musculoskeletal Disorders* 2008;9 doi: 10.1186/1471-2474-9-89

18. Gouttebarge V, Wind H, Kuijer PP, et al. Reliability and Agreement of 5 Ergo-Kit Functional Capacity Evaluation Lifting Tests in Subjects With Low Back Pain. *Archives of Physical Medicine and Rehabilitation* 2006;87(10):1365-70.

19. Grelat M, Gouteron A, Casillas JM, et al. Walking Speed as an Alternative Measure of Functional Status in Patients with Lumbar Spinal Stenosis. *World Neurosurgery* 2019;122:e591-e97.

20. Gulbahar S, Berk H, Pehlivan E, et al. [The relationship between objective and subjective evaluation criteria in lumbar spinal stenosis]. *Acta Orthopaedica et Traumatologica Turcica* 2006;40(2):111-6.

21. Jakobsson M, Brisby H, Gutke A, et al. One-minute stair climbing, 50-foot walk, and timed up-and-go were responsive measures for patients with chronic low back pain undergoing lumbar fusion surgery. *BMC Musculoskeletal Disorders* 2019;20 (1) (no pagination)(137)

22. Kahraman BO, Sengul YS, Kahraman T, et al. Developing a reliable core stability assessment battery for patients with non-specific low back pain. *Spine* 2016;41(14):E844-E50.

23. Kahraman T, Ozcan Kahraman B, Salik Sengul Y, et al. Assessment of sit-to-stand movement in non-specific low back pain: a comparison study for psychometric properties of field-based and laboratory-based methods. *International journal of rehabilitation*

research 2016;Internationale Zeitschrift fur Rehabilitationsforschung. Revue internationale de recherches de readaptation. 39(2):165-70.

24. Lin SI, Lin RM. Disability and walking capacity in patients with lumbar spinal stenosis: Association with sensorimotor function, balance, and functional performance. *Journal of Orthopaedic & Sports Physical Therapy* 2005;35(4):220-26. doi: 10.2519/jospt.2005.35.4.220
25. Lygren H, Dragesund T, Joensen J, et al. Test-retest reliability of the Progressive Isoinertial Lifting Evaluation (PILE). *Spine (03622436)* 2005;30(9):1070-74.
26. Lüder S, Pfingsten M, Lüdtke K, et al. Kann die Aktivitätskapazität von Patienten mit Rückenschmerzen objektiv und reliabel gemessen werden? *physioscience* 2006;2(04):147-55.
27. Maras G, Citaker S, Meray J. Cross-Cultural Adaptation, Validity, and Reliability Study of the Turkish Version of Back Performance Scale. *Spine* 2019;44(1):E39-E44.
28. Maribo T, Iversen E, Andersen NT, et al. Intra-observer and interobserver reliability of One Leg Stand Test as a measure of postural balance in low back pain patients. *International Musculoskeletal Medicine* 2009;31(4):172-77. doi: 10.1179/175361409X12472218841040
29. Maribo T, Schiottz-Christensen B, Jensen LD, et al. Postural balance in low back pain patients: criterion-related validity of centre of pressure assessed on a portable force platform. *European Spine Journal* 2012;21(3):425-31.
30. Moradi B, Benedetti J, Zahlten-Hinguranage A, et al. The value of physical performance tests for predicting therapy outcome in patients with subacute low back pain: a prospective cohort study. *European Spine Journal* 2009;18(7):1041-9.

31. Ocarino JM, Goncalves GGP, Vaz DV, et al. Correlation between a functional performance questionnaire and physical capability tests among patients with low back pain. *Revista Brasileira De Fisioterapia* 2009;13(4):343-49. doi: 10.1590/s1413-35552009005000046

32. Odebiyi DO, Kujero S, Lawal T. Relationship between spinal mobility, physical performance, pain intensity and functional disability in patients with chronic low back pain. *Nigerian Journal of Medical Rehabilitation* 2006

33. Paatelma M, Karvonen E, Heinonen A. Inter- and intra-tester reliability of selected clinical tests in examining patients with early phase lumbar spine and sacroiliac joint pain and dysfunction. *Advances in Physiotherapy* 2010;12(2):74-80.

34. Pfingsten M, Lueder S, Luedtke K, et al. Significance of physical performance tests for patients with low back pain. *Pain Medicine* 2014;15(7):1211-21.

35. Pozo-Cruz Bd, Triviño-Amigo N, Adsuar-Sala JC, et al. Relative and absolute reliability of the progressive iso-inertial lifting test in patients affected by non-specific, chronic low back pain: a 12-week test-retest study. *Rehabilitacion* 2012;46(4):271-76.

36. Rainville J, Childs LA, Pena EB, et al. Quantification of walking ability in subjects with neurogenic claudication from lumbar spinal stenosis--a comparative study. *Spine Journal: Official Journal of the North American Spine Society* 2012;12(2):101-9.

37. Reneman MF, Jorritsma W, Schellekens JMH, et al. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic non-specific low back pain. *Journal of Occupational Rehabilitation* 2002;12(3):119-29.

38. Ryan CG, Gray H, Newton M, et al. The convergent validity of free-living physical activity monitoring as an outcome measure of functional ability in people with chronic low back pain. *Journal of Back and Musculoskeletal Rehabilitation* 2008;21(2):137-42.

39. Smeets RJEM, Hijdra HJM, Kester ADM, et al. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. *Clinical Rehabilitation* 2006;20(11):989-98.

40. Smith RL. Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation. *Journal of Orthopaedic & Sports Physical Therapy* 1994;19(5):277-81.

41. Staartjes VE, Schroder ML. The five-repetition sit-to-stand test: evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine. *Journal of Neurosurgery Spine* 2018;29(4):380-87.

42. Strand LI, Anderson B, Lygren H, et al. responsiveness to change of 10 physical tests used for patients with back pain. *Physical Therapy* 2011;91(3):404-15.

43. Taylor S, Frost H, Taylor A, et al. Reliability and responsiveness of the shuttle walking test in patients with chronic low back pain. *Physiotherapy Research International* 2001;6(3):170-8.

44. Tomkins-Lane CC, Battie MC. Validity and reproducibility of self-report measures of walking capacity in lumbar spinal stenosis. *Spine* 2010;35(23):2097-102.

45. Tomkins-Lane CC, Battie MC, Macedo LG. Longitudinal construct validity and responsiveness of measures of walking capacity in individuals with lumbar spinal stenosis. *Spine Journal: Official Journal of the North American Spine Society* 2014;14(9):1936-43.

46. Tomkins CC, Battie MC, Rogers T, et al. A criterion measure of walking capacity in lumbar spinal stenosis and its comparison with a treadmill protocol. *Spine* 2009;34(22):2444-9.

47. Wand BM, Chiffelle LA, O'Connell NE, et al. Self-reported assessment of disability and performance-based assessment of disability are influenced by different patient characteristics in acute low back pain. *European Spine Journal* 2010;19(4):633-40.

48. Wittink H, Rogers W, Sukiennik A, et al. Physical functioning: self-report and performance measures are related but distinct. *Spine* 2003;28(20):2407-13.

49. Reneman MF, Fokkens AS, Dijkstra PU, et al. Testing lifting capacity: validity of determining effort level by means of observation. *Spine* 2005;30(2):E40-6.

50. Organization WH. International Classification of Functioning, Disability and Health (ICF) 2003 [updated 2 March 2018. Available from: https://www.who.int/classifications/icf/en/.

51. Denteneer L, Van Daele U, Truijen S, et al. reliability of physical functioning tests in patients with low back pain: a systematic review. *Spine Journal* 2018;18(1):190-207.

52. Bassett Jr DR. Validity and reliability issues in objective monitoring of physical activity. *Research quarterly for exercise and sport* 2000;71(sup2):30-36.

53. Leenders NY, Sherman WM, Nagaraja HN. Energy expenditure estimated by accelerometry and doubly labeled water: do they agree? *Medicine and science in sports and exercise* 2006;38(12):2165-72.

54. de Vet HC, Terwee CB, Mokkink LB, et al. Measurement in medicine: a practical guide: Cambridge University Press 2011.

55. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 2016;15(2):155-63.