

# **Validating assessment of spinal stiffness: bench-top performance of the VerteTrack system**

A thesis presented for the degree of Master of Research

Anika Young BChiroSc, MChiroprac

Department of Chiropractic

Faculty of Science and Engineering

Macquarie University

26<sup>th</sup> October 2018

# Table of Contents

<b>SUPERVISORS' STATEMENT .....</b>	<b>V</b>
<b>STATEMENT OF ORIGINALITY .....</b>	<b>VI</b>
<b>ACKNOWLEDGMENTS .....</b>	<b>VII</b>
<b>SCHOLARSHIP FUNDING .....</b>	<b>VIII</b>
<b>CONFLICT OF INTEREST STATEMENT.....</b>	<b>VIII</b>
<b>LIST OF ABBREVIATIONS.....</b>	<b>IX</b>
<b>ABSTRACT .....</b>	<b>X</b>
<b>INTRODUCTION.....</b>	<b>1</b>
1.1 THE GLOBAL BURDEN OF NON-SPECIFIC LOW BACK PAIN .....	1
1.2 MANUAL THERAPY IN THE MANAGEMENT OF SPINAL PAIN .....	1
1.3 CLINICAL ASSESSMENT OF SPINAL STIFFNESS.....	2
1.4 SYSTEMS EMPLOYED TO ASSESS SPINAL STIFFNESS .....	3
1.4.1 <i>Manual assessment of spinal stiffness</i> .....	3
1.4.2 <i>Mechanical assessment of spinal stiffness</i> .....	4
1.5 THE VERTETRACK: A NEW DEVICE FOR SPINAL STIFFNESS ASSESSMENT.....	7
1.5.1 <i>Concepts for evaluation of a novel mechanical device</i> .....	8
1.6 THESIS AIMS .....	9
1.7. HYPOTHESES .....	9
1.8 THESIS OVERVIEW .....	9
<b>2. SYSTEMATIC LITERATURE REVIEW.....</b>	<b>10</b>
2.1 REVIEW RATIONALE .....	10
2.2 DATA SOURCE AND SEARCH STRATEGY .....	10
2.3 ELIGIBILITY CRITERIA .....	11
2.4 DATA EXTRACTION AND CRITICAL APPRAISAL .....	11
2.5 DATA SYNTHESIS.....	12
2.6 STUDIES IDENTIFIED AND INCLUDED .....	13
2.7 STUDY CHARACTERISTICS .....	13
2.7.1 <i>Critical appraisal tools used by included reviews</i> .....	14
2.8 CRITICAL APPRAISAL OF INCLUDED REVIEWS .....	17
2.9 MANUAL SPINAL STIFFNESS ASSESSMENT: INTER AND INTRA-EXAMINER RELIABILITY .....	18
2.10 MECHANICAL DEVICES: TEST-RETEST RELIABILITY.....	18
2.11 SUMMARY OF PRINCIPAL FINDINGS .....	19

2.12 SUMMARY OF METHODOLOGICAL ISSUES IN THE FIELD .....	19
2.13 GENERALISABILITY.....	20
2.14 CHAPTER 2 SUMMARY .....	20
<b>3. METHODS: BENCH-TOP PERFORMANCE OF THE VERTETRACK SYSTEM.....</b>	<b>21</b>
3.1 CONTEXT.....	21
3.2 STUDY DESIGN, SETTING AND EQUIPMENT .....	21
3.2.1 <i>The VerteTrack</i> .....	22
3.2.2 <i>VerteTrack custom software</i> .....	23
3.2.3 <i>Indentation procedure</i> .....	24
3.3 EXPERIMENT ONE: PRECISION AND BIAS OF THE VERTETRACK FOR THE APPLIED LOAD .....	25
3.3.1 <i>Precision of load</i> .....	25
3.3.2 <i>Bias of load</i> .....	26
3.4 EXPERIMENT TWO: PRECISION AND BIAS OF THE VERTETRACK INDENTER HEAD DISPLACEMENT .....	27
3.4.1 <i>Precision of VerteTrack RIH displacement</i> .....	27
3.4.2 <i>Bias of VerteTrack RIH displacement</i> .....	27
3.5 EXPERIMENT THREE: PERFORMANCE OF THE VERTETRACK SYSTEM DURING BOTH STATIC AND DYNAMIC MODES OF OPERATION .....	28
3.5.1 <i>Methods of indentation and medium characteristics</i> .....	28
<i>Static indentation</i> .....	29
<i>Dynamic indentation</i> .....	29
3.5.2 <i>Precision of stiffness</i> .....	30
3.5.3 <i>Bias of stiffness</i> .....	30
3.6 PROJECT MANAGEMENT .....	31
<b>4. RESULTS.....</b>	<b>32</b>
4.1 EXPERIMENT ONE: PRECISION AND BIAS OF THE VERTETRACK FOR THE APPLIED LOAD .....	32
4.1.1 <i>Precision of load</i> .....	32
4.1.2 <i>Bias of load</i> .....	32
4.2 EXPERIMENT TWO: PRECISION AND BIAS OF THE VERTETRACK INDENTER HEAD DISPLACEMENT .....	35
4.2.1 <i>Precision of RIH displacement</i> .....	35
4.2.2 <i>Bias of RIH displacement</i> .....	35
4.3 EXPERIMENT THREE: PERFORMANCE OF THE VERTETRACK SYSTEM DURING BOTH STATIC AND DYNAMIC MODES OF OPERATION .....	38
4.3.1 <i>Precision of stiffness</i> .....	38
4.3.2 <i>Agreement of VerteTrack Static and Dynamic stiffness measurements</i> .....	38
<b>5. DISCUSSION.....</b>	<b>41</b>
5.1 OVERVIEW OF MAIN FINDINGS .....	41

5.1.1 Reliability of existing spinal stiffness assessment methods .....	41
5.1.2 Bench-top performance of the VerteTrack.....	42
5.2 STRENGTHS .....	43
5.2.1 Reliability of existing spinal stiffness assessment methods .....	43
5.2.2 Bench-top performance of the VerteTrack.....	44
5.3 LIMITATIONS .....	44
5.3.1 Reliability of existing spinal stiffness assessment methods .....	44
5.3.2 Bench-top performance of the VerteTrack.....	45
5.4 FUTURE DIRECTIONS .....	45
5.5 CONCLUDING STATEMENT .....	46
<b>REFERENCES .....</b>	<b>47</b>
<b>APPENDIX A: SEARCH STRING.....</b>	<b>54</b>
<b>APPENDIX B: AMSTAR-2 DOMAINS .....</b>	<b>55</b>
<b>APPENDIX C. PRIMARY RELIABILITY STUDIES.....</b>	<b>56</b>

## Supervisors' statement

As supervisors of Anika Young, we certify that we consider this thesis "Validating assessment of spinal stiffness: bench-top performance of the VerteTrack system" is sufficiently well presented to be examined and does not exceed the prescribed page limit.

**Aron Downie** BSc, MChir, MPhil

Department of Chiropractic, Faculty of Science and Engineering, Macquarie University



Date: 26<sup>th</sup> October 2018

**Michael Swain** BChiroSc, MChiroprac, MPhil

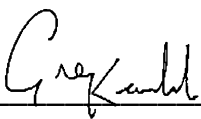
Department of Chiropractic, Faculty of Science and Engineering, Macquarie University



Date: 26<sup>th</sup> October 2018

**Professor Greg Kawchuk** BSc, DC, MSc, PhD

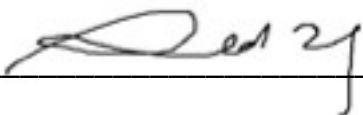
Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta



Date: 26<sup>th</sup> October 2018

**Doctor Arnold Wong** BSc, BScPT, MPhil, PhD

Department of Rehabilitation Sciences, Faculty of Health and Social Sciences, The Hong Kong Polytechnic University



Date: 26<sup>th</sup> October 2018

## **Statement of Originality**

I certify that this work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due acknowledgement is made itself in text.

*A Young* Date: 26<sup>th</sup> October 2018

## **Acknowledgments**

I would like to thank all my supervisors Aron, Mike, Greg and Arnold for your patience, support and for showing me the ropes. Your expertise has helped me to develop a deeper understanding of the complexities and multifactorial nature of spinal stiffness research. Your feedback and opinions have caused me to challenge my preconceived notions and enabled me grow as a researcher and as a clinician.

A special thanks to Aron, I do not know what I would have done without your mentorship, enthusiasm, patience, tech support and assistance trouble-shooting challenges. You have really gone above and beyond to ensure my success. Your ongoing support and assistance is greatly appreciated.

## **Scholarship funding**

I would like to acknowledge the generous scholarship support (MRes scholarship) I have received from the Australian Chiropractors' Association (ACA). The ACA was not involved in the collection, analysis or interpretation of data presented in this thesis.

## **Conflict of Interest Statement**

I, Anika Young declare no conflicts of interest, financial or otherwise in completing this thesis. Professor Greg Kawchuk invented the VerteTrack system. Potential conflicts of interest were managed in this thesis. I, Anika Young collected, extracted and analysed the data, reported and interpreted the results. Professor Kawchuk provided high level overview of the machine operation and confirmed the correctness of data output.



## **List of abbreviations**

AD	Aron Downie
AY	Anika Young
F-D curve	Force displacement curve
GK	Greg Kawchuk
HVLA	High-velocity low amplitude
IFD	Indentation force deflection
LBP	Low back pain
M	Mechanical
MA	Mechanically assisted
MCID	Minimally clinically importance difference
mm	millimetre
MSSA	Manual spinal stiffness assessment
N	Newton
P-A	Posterior to anterior
PAIVM	Passive accessory intervertebral motion
PICO	Population, intervention, comparator and outcome
RIH	Rolling indenter head
RoB	Risk of bias
SMT	Spinal manipulative therapy

## **Abstract**

In the conservative management of non-specific spinal pain, manual therapists commonly assess spinal stiffness to identify aberrant joint motion and to direct treatment. There are various manual and mechanical test methods employed to assess spinal stiffness. The validation of spinal stiffness assessment methods is a multistep process. This thesis has two discrete objectives: (i) to review the literature on reliability of manual and mechanical methods used in the assessment of spinal stiffness, and (ii) determine the bench-top performance of a novel mechanical spinal stiffness assessment device, the VerteTrack. The VerteTrack was designed to measure spinal stiffness at either a single spinal level (static indentation), or multiple spinal levels (dynamic indentation). A review of the literature found that for the assessment of spinal stiffness, manual methods had variable reliability whereas mechanical methods had high reliability but limited clinical utility. The bench-top performance of the VerteTrack demonstrated a high level of accuracy (equivalent to the resolution of the reference test equipment). In this study, comparison of dynamic to static indentation modes revealed a small negative systematic bias (lower spinal stiffness). Future research is required to assess the reliability of the VerteTrack in human subjects.

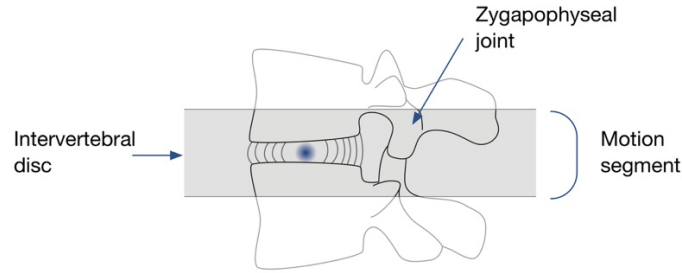
# Introduction

## 1.1 The global burden of non-specific low back pain

Low back pain (LBP) has increased significantly over the past 25 years and is now the leading cause of disability globally<sup>1,2</sup>. In monetary terms, the total costs of LBP in Australia (2001) exceeds \$9billion<sup>3</sup>. For individuals who experience LBP, recurrence is common and recovery is often incomplete<sup>4,5</sup>. In 10% of cases, LBP is a chronic/persistent condition much like diabetes and cardiovascular disease<sup>6</sup>. Low back pain is defined as pain located between the lower border of the 12<sup>th</sup> rib and the gluteal fold<sup>7</sup>. Only 10% of LBP cases have identifiable tissue damage or a pathophysiological cause. The remaining 90% of LBP cases have no identifiable tissue damage and are referred to as non-specific LBP<sup>7</sup>. Given the absence of a pathological aetiology, clinical management of non-specific LBP frequently targets the management of pain<sup>4</sup>. Common goals of non-specific LBP management include the reduction of pain and functional limitations. It typically involves multimodal interventions including education, advice, pharmacological and/or non-pharmacological therapies (such as manual therapies, heat/ice, and exercise)<sup>7</sup>.

## 1.2 Manual therapy in the management of spinal pain

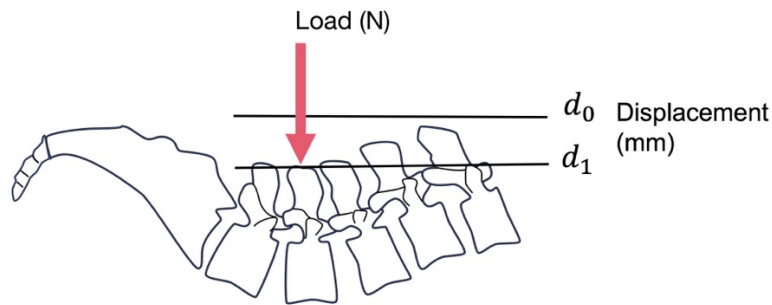
A subset of non-pharmacological management options includes manual therapy. Commonly employed manual therapy interventions include, but are not limited to, spinal manipulative therapy (SMT) and mobilisation. High-velocity low amplitude (HVLA) SMT is one type of manual therapy where a thrust force is applied to a spinal motion segment (Figure 1.1) at the end of joint range. The application of HVLA-SMT often creates increased joint separation resulting in a fluid cavitation (audible ‘crack’ sound)<sup>8</sup>. Therapeutic rationales that underpin the application of SMT or mobilisation relate to neuro-musculoskeletal effects including the transient increase in lumbar segmental motion and hypoalgesia (reduction in pain)<sup>9,10</sup>. Clinical practice guidelines recommend SMT in conjunction with other treatment modalities in the management of patients with acute and chronic LBP<sup>11</sup>.



**Figure 1.1** Lumbar spine motion segment

### 1.3 Clinical assessment of spinal stiffness

Manual therapists commonly use physical tests to detect spinal hypomobility which is often the same location for SMT application<sup>12</sup>. Spinal hypomobility can be conceptualised as a reduction in accessory motion at the level of the intervertebral disc and facet joints (motion segment) known as passive accessory intervertebral motion (PAIVM)<sup>13</sup>. Spinal hypomobility is quantified by measuring spinal stiffness. Spinal stiffness is a measurement of spinal translation, defined as the displacement of spinal and paraspinal tissues due to the application of a posterior to anterior (P-A) load at a spinal segment (Figure 1.2)<sup>14</sup>. In simple terms, spinal stiffness is a reduction in P-A translation of one vertebra, relative to adjacent vertebrae at a given load.



$$\text{Stiffness} = \frac{\text{Load (N)}}{\text{Displacement (mm)}}$$

*Equation 1.1*

**Figure 1.2** Calculation of spinal stiffness (N/mm) is achieved by measuring displacement of spinal tissues ( $d_0 - d_1$ ) due to applied perpendicular load when the patient is in a prone position (Equation 1.1)

## **1.4 Systems employed to assess spinal stiffness**

For the assessment of spinal stiffness in human subjects, different test systems have been employed. For example, a simple (qualitative) system may comprise a clinician-applied load to a spinal motion segment, with estimation of tissue displacement and thus spinal stiffness (manual spinal stiffness assessment, MSSA). An example of a more complex (quantitative) test system may comprise a series of machine applied loads to a spinal motion segment, with automated measurement of tissue displacement for calculation of spinal stiffness using custom software (mechanical assessment).

To be clinically useful, a spinal stiffness test system must be reliable under a range of conditions. Three main types of reliability are pertinent when assessing the spinal stiffness of human subjects<sup>15</sup>. First, intra-rater reliability relates to an examiner's ability to reproduce the same findings upon repeated testing under the same conditions. Second, inter-rater reliability refers to two or more independent examiners being able to reproduce the same findings. Third, test-retest reliability refers to the stability of the test system over time. Chapter two reviews the literature relating to the reliability of different test systems used to assess spinal stiffness of human subjects.

### 1.4.1 Manual assessment of spinal stiffness

In clinical practice, spinal hypomobility is commonly determined by a clinician during the physical examination. Manual spinal stiffness assessment is a frequently used clinical test for both diagnoses and as an outcome measure<sup>16</sup>. Manual spinal stiffness assessment is typically performed by hand with a clinician applying a P-A load to the patient's spinous process where the resultant movement is then qualitatively interpreted<sup>13</sup>. Results of MSSA are typically reported as "hypomobile" (decreased movement), "normal", or "hypermobile" (increased spinal movement) which provides clinical inference towards potential kinesiotherapy<sup>10,17</sup>. Despite being common in practice, there are concerns about the reliability and validity of this physical test given the qualitative and subjective interpretation of MSSA<sup>16</sup>.

The reliability of MSSA reported in the literature exhibits a large degree of heterogeneity. Variation in reliability may arise from the use of heterogeneous study samples (e.g. combined symptomatic and asymptomatic populations), patient positioning (prone, supine or seated),

contacts applied by the clinician (thumb or pisiform) and rating scale (ranges from dichotomous to an eleven-point scale)<sup>14,17,18</sup>. Additionally, factors related to the execution of the test have been shown to influence the results of MSSA<sup>17</sup>. For example, variability in clinician contact force, speed and angle may influence the perceived spinal stiffness<sup>17,19,20</sup>. Given the heterogeneity of available evidence, the veracity of conclusions for the reliability of MSSA remains unclear.

A challenge for researchers in the field has been the absence of an accepted reference standard for the assessment of spinal stiffness<sup>21</sup>. In part, this has led to the development of mechanical assessment systems which provide quantitative (objective) measurement of stiffness. The few studies that have compared MSSA to mechanical assessment have found poor correlation and thus low criterion validity for MSSA<sup>22-24</sup>.

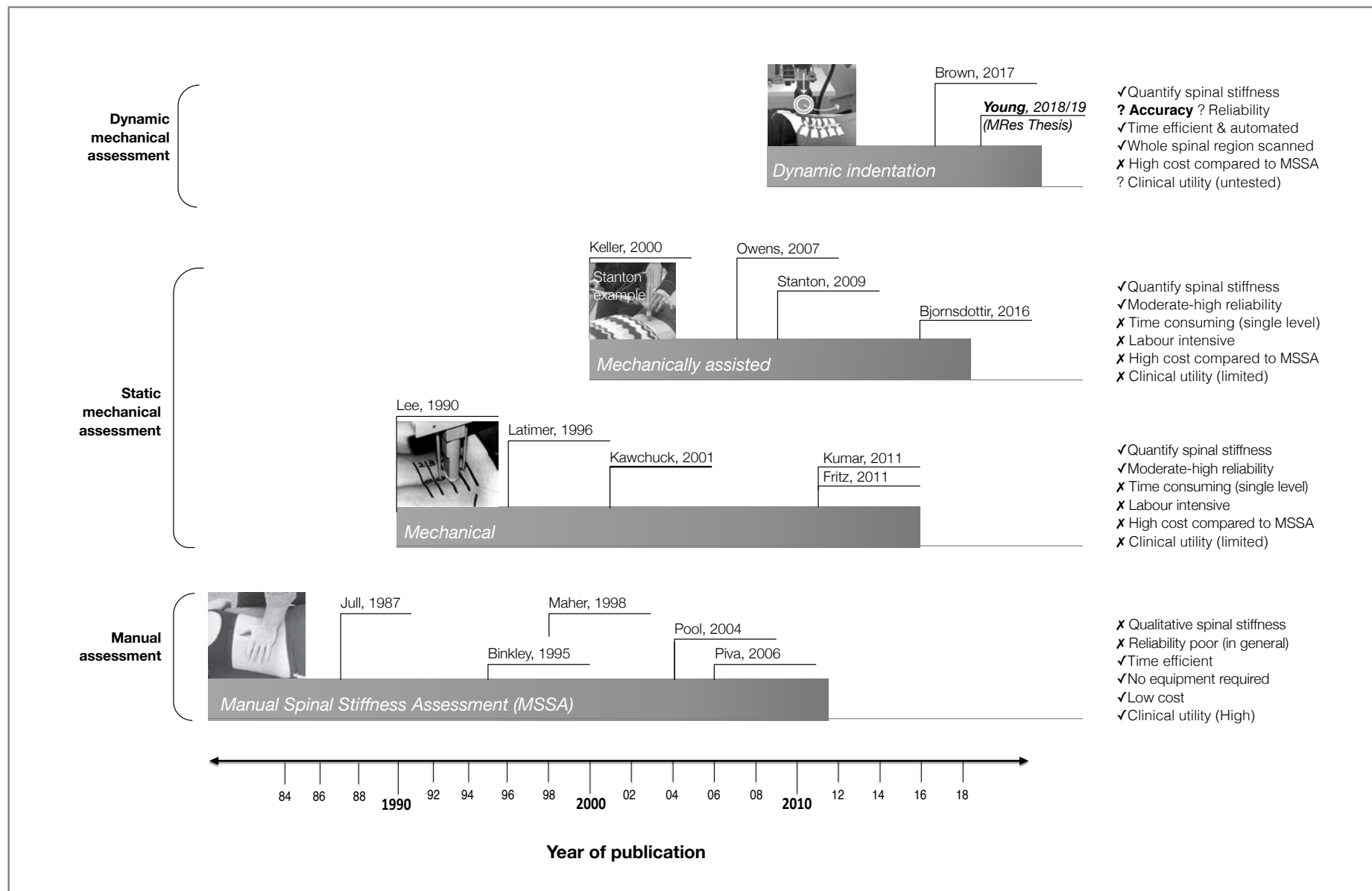
#### 1.4.2 Mechanical assessment of spinal stiffness

Typically, a mechanical device used for the assessment of spinal stiffness aims to (i) constrain variability during the procedure (e.g. by use of repeatable contact force, speed and angle), and (ii) quantify stiffness using a standardised measure that is not dependent on clinician judgment. For instance, numerous qualitative scales used in MSSA can be replaced with a single quantitative measure of stiffness, reported as Newtons per millimetre (N/mm). The reliability across different mechanical designs has demonstrated good test-retest reliability, meaning that devices are able to reproduce the same spinal stiffness results repeatedly<sup>21,25-33</sup>.

Whilst multiple devices have been developed to assess spinal stiffness, the basic mechanism of assessment is similar across all devices: a mechanical indenter head contacts the patient's spine to apply a defined load, and the resultant displacement is then recorded (e.g. by P-A translation of the indenter head or tissue displacement visualised using ultrasonography). Two broad categories of devices have been described based on the level of operator involvement in the assessment: (i) "*mechanical*", and (ii) "*mechanically assisted*"<sup>17</sup>. *Mechanical devices* are generally mounted within a frame and require minimal intervention by the operator. In contrast, *mechanically assisted* devices are more portable and require direct assistance from the operator to perform an assessment. For both device categories, the patient lies prone on a testing plinth whilst the indenter head travels in a P-A plane at the

target vertebral level (e.g. L4) with spinal stiffness calculated using Equation 1.1. Using an incremental increase in spinal P-A load, a spinal stiffness “trace” can be graphed, which demonstrates a broadly linear relationship between load and displacement (commonly known as a force-displacement curve)<sup>34</sup>.

The mechanical assessment of spinal stiffness has partially solved the heterogeneity and reliability issues of MSSA but has introduced a new set of challenges regarding the feasibility and integration of mechanical devices into both laboratory and clinical settings. There has been an evolution of devices to address the limitations of mechanical and mechanically assisted assessment of spinal stiffness (Figure 1.3). However, current devices are mostly expensive, time-consuming and compromise patient comfort<sup>35</sup>.



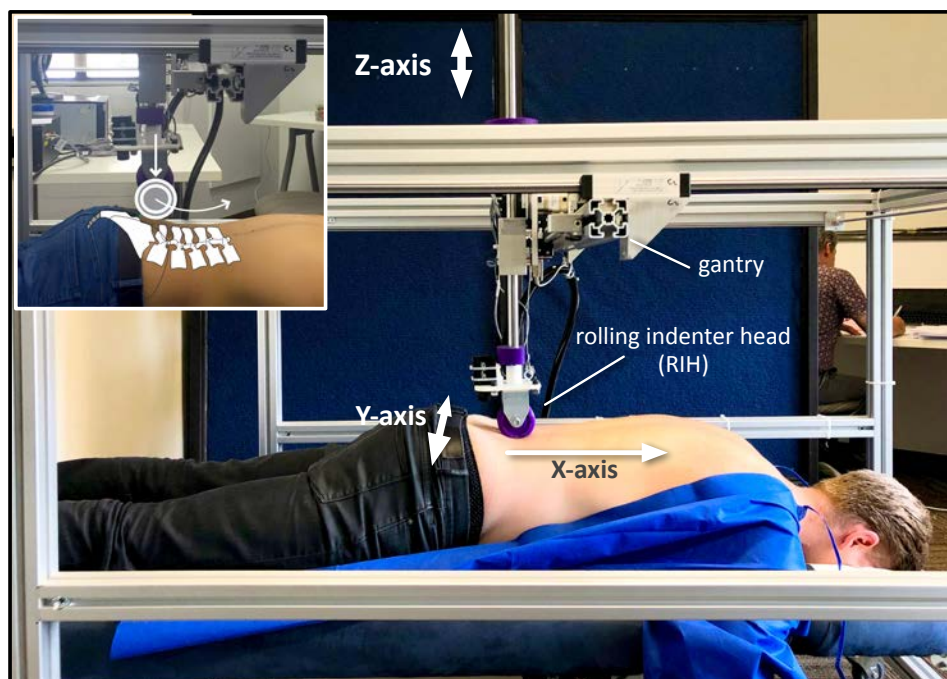
**Figure 1.3** The evolution of test systems used to assess spinal stiffness



## 1.5 The VerteTrack: A new device for spinal stiffness assessment

In response to the logistical limitations of existing mechanical devices (specifically in reference to expense, assessment duration and comfort), Professor Greg Kawchuk of the University of Alberta has developed a novel device to assess spinal stiffness called the VerteTrack<sup>35</sup>. The VerteTrack is a mechanical device designed to collect stiffness data from multiple vertebral levels per indentation sequence, compared to other mechanical devices which assess stiffness at a single vertebral level.

The VerteTrack consists of roller wheels attached to an indenter head mounted on an aluminium gantry. The indenter head has triaxial movement: X-axis (longitudinal, inferior–superior), Y-axis (transverse, left–right) and Z-axis (vertical, posterior–anterior). During spinal stiffness assessment, VerteTrack applies discrete incremental loads (a series of plates with the nominal mass of 1kg) along the Z-axis. The use of a series of plates instead of a programmed force transducer increases design simplicity and minimises system cost. Force is applied to a subject's spine via a rolling indenter head (RIH) which straddles a subject's spinous process by contacting paraspinal soft tissues. The design of the rolling indenter head increases the contact surface area and does not apply force directly to bony prominences which consequently increases patient comfort<sup>35</sup>.

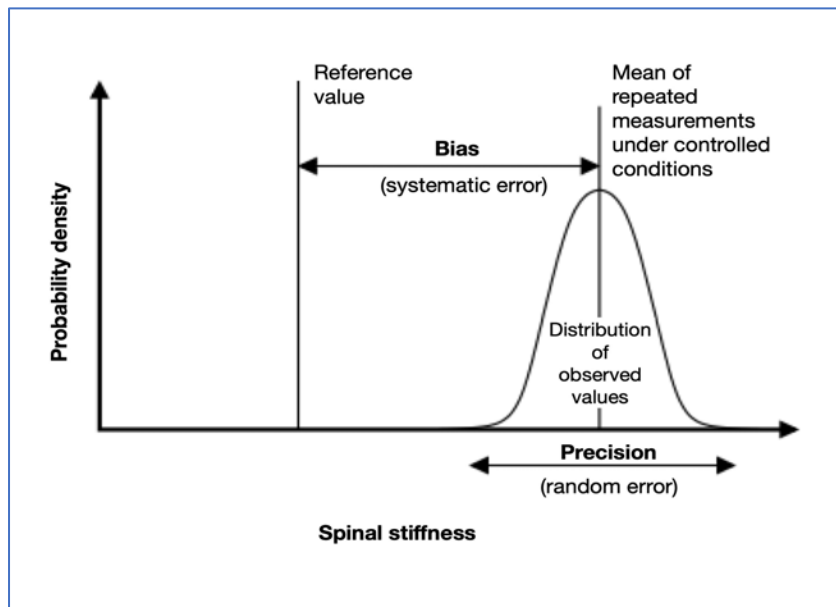


**Figure 1.4** The VerteTrack positioned over a patient simulating dynamic indentation

The VerteTrack has the option to perform both static and dynamic indentation. Static indentation assesses a single spinal level. During static indentation, there is only movement about the Z-axis with no concurrent movement about X or Y-axes. In contrast, the innovation of dynamic indentation allows the indenter head to follow a pre-set trajectory about the X and Y-axes to include multiple segments, while concurrently applying a P-A load following the sagittal curvature of the lumbar spine. Moreover, dynamic assessment has been shown to reduce assessment without compromise to patient comfort<sup>35</sup>.

### 1.5.1 Concepts for evaluation of a novel mechanical device

The test method delivered by the VerteTrack has not been evaluated for performance in a controlled (bench-top) environment<sup>36,37</sup>. Accuracy is a key property to be evaluated during bench-top assessment and is influenced by both systematic and random errors<sup>36</sup>. Therefore, the bench-top performance of the VerteTrack can be quantified by the observational errors from repeated measurements under controlled conditions and is a combination of the systematic error (bias) and the random error (precision). The magnitude of both systematic and random errors is necessary to describe the overall accuracy of the VerteTrack<sup>36,38</sup> (Figure 1.5). Chapters 3 and 4 explore testing of the VerteTrack system.



**Figure 1.5** Terminology used to describe the bench-top performance of a test system to measure spinal stiffness using repeated measurements under controlled conditions. The magnitude of both systematic and random error are necessary to describe the overall accuracy of the test system<sup>36,38</sup>

## **1.6 Thesis aims**

This thesis aims to examine issues pertinent to the reliability and accuracy of spinal stiffness assessment. Specifically, the thesis aims are to:

1. Review the published literature regarding the reliability of manual and mechanical test systems used for the assessment of spinal stiffness.
2. Investigate the bench-top performance of the VerteTrack system for the assessment of spinal stiffness.

## **1.7. Hypotheses**

In relation to the second aim, it is hypothesised that:

1. The bench-top performance of the VerteTrack system exhibits appropriate levels of accuracy for future validation testing using human subjects.
2. That the bench-top performance of the VerteTrack system is equivalent during both static and dynamic modes of operation.

## **1.8 Thesis overview**

This dissertation contains two linked projects that relate to the central theme of the thesis which is “The reliability of spinal stiffness assessment: the bench-top performance of the VerteTrack system”. The thesis includes an umbrella systematic literature review that evaluates the reliability of manual and mechanical assessment of spinal stiffness (Chapter 2), and a study that investigates the benchtop performance of the VerteTrack system for measuring spinal stiffness (Chapters 3 and 4). The studies are presented in a thesis format as per the requirements of the degree (Master of Research).

## 2. Systematic literature review

### 2.1 Review rationale

Several systematic reviews have evaluated the reliability of test systems used to assess spinal stiffness<sup>14,18,39-44</sup>. Syntheses to date have been limited by superficial review methodologies and narrow-focused study samples. No single systematic review has comprehensively explored the topic or accounted for the rapid evolution of technologies developed to assess spinal stiffness. An umbrella review provides a framework to synthesise evidence from existing reviews and to compare the recommendations for the assessment of spinal stiffness<sup>45</sup>. An umbrella systematic literature review was conducted<sup>45</sup>. The study protocol was pre-specified, and the review was conducted in accordance with PRISMA guidelines<sup>46</sup>.

This study aimed to (i) provide a high-level summary for the intra-rater, inter-rater and test-retest reliability (Table 2.1) of manual and mechanical methods used to measure spinal stiffness in the cervical, thoracic and lumbar spine; (ii) compare and contrast the findings from published systematic reviews; and (iii) identify any knowledge gaps or methodological weaknesses in the field of spinal stiffness assessment.

**Table 2.1** Reliability Terminology<sup>47</sup>

Term	Definition
Intra-rater reliability	A measure of agreement between repeated measurements by one rater.
Inter-rater reliability	A measure of agreement between 2 or more raters measuring the same subject/s.
Test-retest reliability	A measure of agreement between repeated measurements of the same subject under the same conditions using the same instrument.

### 2.2 Data source and search strategy

An electronic search of databases PubMed, Embase and Cinahl was conducted for reviews published between January 2000 and April 2018. A Restriction was placed on the earliest publication date due to the likelihood that reviews published prior to 2000 would have been updated over the following 18 years, or would be superseded if not. Search terms relating to keywords spinal stiffness, motion palpation, spinal palpation, intervertebral motion, intersegmental motion, posteroanterior stiffness, lumbar stiffness, thoracic stiffness and

cervical stiffness were used (Appendix A: search string). A search filter was applied to the search to include only review study designs. The electronic search was supplemented with hand searches and backward citation tracking, performed using the included articles reference list. One rater (AY) performed title screening and abstract screening. Two raters (AY, AD) independently assessed full-text articles for eligibility. Any discrepancies were discussed and resolved by consensus.

## **2.3 Eligibility criteria**

Eligible studies were systematic, scoping or narrative reviews. Studies were eligible if they quantified the reliability of manual and/or mechanical assessment methods of spinal stiffness for both symptomatic and asymptomatic human populations. For the purposes of this review, *manual assessment* was defined as performed by a clinician without mechanical assistance. Likewise, *mechanical assessment* was defined as a clinician or researcher assisted by a mechanical device. Studies were excluded: if they exclusively assessed peripheral joint systems, included participants with pathology, studied the effect of an intervention as a primary aim, or were not written in English.

## **2.4 Data extraction and critical appraisal**

Data were independently extracted from included reviews by two raters (AY, AD). Any disagreements were resolved by consensus. Data were extracted using a standardised method for umbrella reviews<sup>45</sup>. Where available, data were extracted from each study relating to year of publication, objective of the review, spinal region(s) examined, method(s) of assessment, symptomatic/asymptomatic population, number of primary studies, instrument used for quality appraisal of primary studies, quantitative analysis methodology, and study outcomes used to assess reliability of spinal stiffness. Where feasible, reliability data were also extracted from each primary study to inform discussion relating to identified knowledge gaps or methodological weaknesses in the field of spinal stiffness assessment.

Two raters (AY, AD) independently appraised each included review for quality. A modified form of the AMSTAR-2 tool was used to critically appraise the literature, given that a validated instrument designed for this study's aims does not exist<sup>48</sup>. The AMSTAR-2 tool comprises 16 domains related to 1) PICO (population, intervention, comparator and

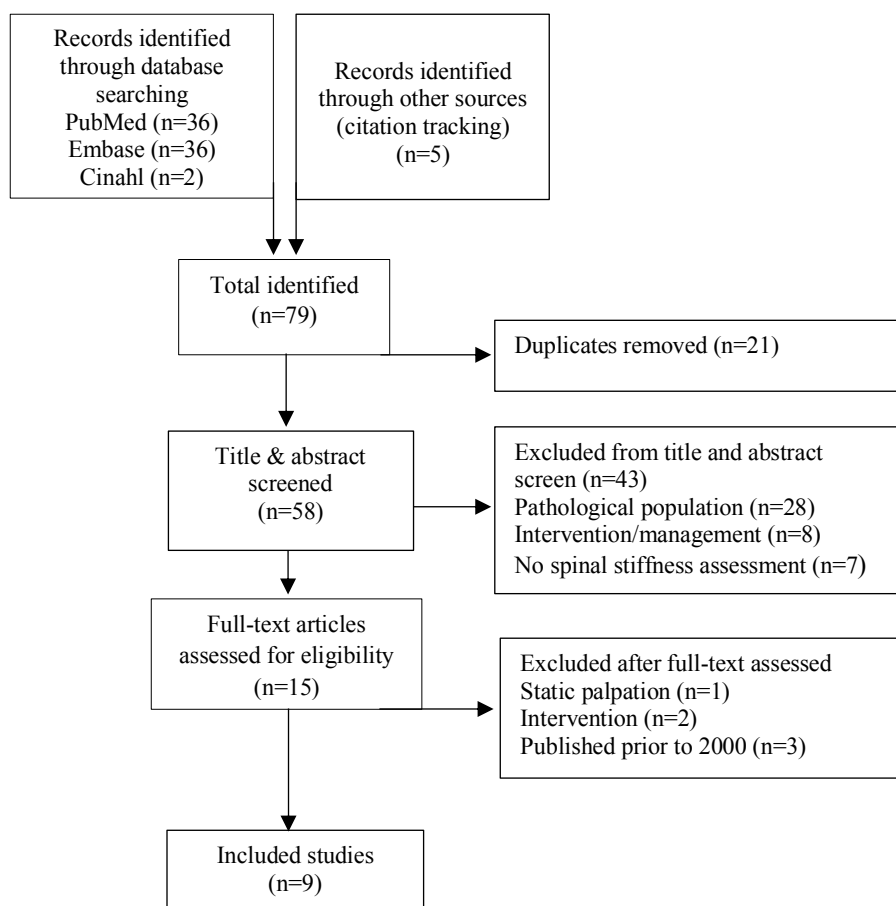
outcome), 2) registered protocol prior to review commencement, 3) description of included study designs, 4) detailed search strategy, 5) study selection was performed in duplicate, 6) data extracted in duplicate, 7) list of excluded studies, 8) description of the included studies, 9) assessment of risk of bias (RoB), 10) funding disclosed, 11) if meta-analysis was performed appropriate methods were used, 12) if meta-analysis was performed RoB was factored in, 13) authors factored in RoB into discussion, 14) description of heterogeneity, 15) if meta-analysis was performed discussion of publication bias and 16) conflicts of interest report (Appendix B: AMSTAR-2 domains)<sup>48</sup>. The tool scores each item as *yes*, *partial yes* or *no*. For this review, modification to the AMSTAR-2 comprised omission of the summary score, given that domains used to calculate the summary score favoured interventional study designs. There was no other modification to the AMSTAR-2 tool.

## 2.5 Data synthesis

Reliability data were extracted from each review and reported per spinal region. To account for some anatomical overlap in studies that assessed multiple spinal levels, the following categorisation was used: (1) *cervical region*, studies which predominantly investigated levels between C1–T1; (2) *thoracic region*, between C7–L1; and (3) *lumbar region*, between T12–S1. The various manual or mechanical assessment systems were broadly categorised into (1) *manual spinal stiffness assessment (MSSA)*, where assessment was by hand only; (2) *mechanical devices*, which required no contribution of force by the examiner; and (3) *mechanically assisted devices*, which required assistance from the examiner. Data was reported by spinal region where available. To account for the diversity of statistical methods used to report different types of reliability, both Cohen's Kappa ( $\kappa$ ) and intra-class correlation (ICC) coefficients were extracted were reported. Cohen's Kappa were rated according to Landis and Koch (1977) ( $\kappa$ : *poor* < 0.20, *fair* 0.21–0.40, *moderate* 0.41–0.60, *good* 0.61–0.80 and *very good* 0.91–1.0)<sup>49</sup>. Intraclass correlation coefficients were interpreted according to Rankin and Stokes (1998) (ICC; *poor* 0–0.40, *fair to good* 0.41–0.75 and *good* 0.75–1.00)<sup>50</sup>.

## 2.6 Studies identified and included

Electronic database searches (PubMed, Embase and Cinahl) identified 53 unique studies of interest (Figure 2.1). After the screening of titles and abstracts, 15 full-text reviews were retrieved. A total of nine studies met the criteria for inclusion. Included studies were Haneline et al. (2008)<sup>39</sup>, Hollerwoger (2006)<sup>40</sup>, Huijbregts (2002)<sup>41</sup>, Jonsson et al. (2018)<sup>42</sup>, Seffinger et al. (2004)<sup>18</sup>, Snodgrass et al. (2012)<sup>14</sup>, Stochkendahl et al. (2006)<sup>43</sup>, Van Trijffel et al. (2005)<sup>44</sup> and Wong et al. (2017)<sup>17</sup>.



**Figure 2.1** PRISMA flow

## 2.7 Study Characteristics

The characteristics of the nine included studies are summarised in Table 2.2. Eight reviews (89%) examined both symptomatic and asymptomatic populations. Two reviews (22%) investigated manual and mechanical methods of spinal stiffness assessment<sup>14,17</sup>, the remaining reviews examined manual spinal stiffness assessment only<sup>39-44</sup>. The majority of reviews reported reliability data for all three spinal regions<sup>14,18,39,41,43</sup>, the remainder reporting

either two regions<sup>40,42</sup>, or single spinal region<sup>17</sup>. From the nine included reviews, 74 unique primary reliability studies were identified. The number of primary studies included in each review ranged from 11<sup>42</sup> to 104<sup>14</sup>. The primary reliability studies spanned publication dates 1976<sup>51</sup>–2016<sup>33</sup>.

### 2.7.1 Critical appraisal tools used by included reviews

There was no consistency in the quality assessment tools used by the included reviews to score methodological bias in primary reliability studies. For example, three reviews used a non-validated tool: Stochkendahl et al. (2006)<sup>43</sup> developed their own tool to critically appraise reproducibility studies by including domains for randomisation, case mix (symptomatic/asymptomatic), blinding and statistical reporting<sup>43</sup>; Haneline et al. (2008)<sup>39</sup> used the tool developed by Stochkendahl et al. (2006)<sup>43</sup> and Seffinger et al. (2004)<sup>18</sup> developed a quality appraisal tool based on five domains relating to study subjects, examiners, study conditions, data analysis, and reporting of results. In contrast, three reviews used validated assessment tools: Jonsson et al. (2018)<sup>42</sup> used the Quality Appraisal of Reliability Studies (QAREL) tool; Hollerwoger (2006)<sup>40</sup> used the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool; and Van Trijffel et al. (2005)<sup>44</sup> adapted criteria based on Standards for Reporting of Diagnostic Accuracy (STARD) and QUADAS tools with criteria related to study sample, examiners, standardisation of assessment, intra-examiner reliability, blinding, attrition, appropriate methods for calculating reliability and bias. The remaining three reviews did not critically appraise study quality of primary reliability studies<sup>14,17,41</sup>.

Due to the heterogeneity in quality assessment tools used by included reviews, comparison of recommendations based on primary study quality was not possible. Further, for reviews that used the same tool, there was a lack of consistency between reviews when scoring the same study (Appendix B: AMSTAR-2 domains).



**Table 2.2** Characteristics of included reviews

Citation (Study design)	Study objective(s)	Primary study publication range (Number of studies)*†	Spinal region	Assessment method	Quality appraisal tool used	Reliability findings	Method of synthesis/analysis employed to synthesize the evidence	Comments
Haneline et al. (2008) <sup>39</sup> (Systematic review)	To report the reliability of manual spinal stiffness assessment stratified by spinal region.	1985-2006 (n=44)*†	C°, T° and L°	MSSA	Stochkendahl quality of reproducibility studies	<b>Inter-examiner reliability</b> Cervical (poor–good) Thoracic (poor–good) Lumbar (poor–good) <b>Intra-rater reliability</b> Cervical (poor–very good) Thoracic (moderate–good) Lumbar (poor–moderate)	Data were extracted according to population, number of examiners, examiner discipline, method used to assess and reporting used. A forest plot was used to graphically represent the data. Statistical significance was tested for difference between proportions using Yates-corrected.	Some studies focused on particular regions eg. mid-lower cervical or upper cervical and some assessed reliability across cervical-mid thoracic regions. Reliability was reported in k/k(w) values, r, ICC and PA.
Hollerwoger (2006) <sup>40</sup> (Systematic review)	To report the reliability of cervical spine manual spinal stiffness assessment.	1989-2004 (n=15)*†	C°	MSSA	QUADAS	<b>Inter-rater reliability</b> Cervical (poor–very good)	Data were pooled with k values and PA graphically represented.	No table was presented for primary data. Cervical regions varied with studies; upper, mid and lower cervical spine
Huijbregts (2002) <sup>41</sup> (Systematic review)	To report the reliability of manual spinal stiffness assessment stratified by spinal region.	1982-2000 (n=28)*†	C°, T° and L°	MSSA	-	<b>Inter-rater reliability</b> Cervical (poor–good) Thoracic (poor–moderate) Lumbar (poor–good) <b>Intra-rater reliability</b> Cervical (poor–very good) Thoracic (fair–very good) Lumbar (poor–good)	Data were extracted according to intra- and inter-examiner reliability. Detailed pertaining to sample, number of participants, examiners, method of exam, region and reporting was commented on in the results.	Regions examined in primary studies were varied. Reliability was reported in k/k(w)/k(m) values, ICC, r and qualitative reporting.
Jonsson et al. (2018) <sup>42</sup> (Systematic review)	To report the reliability of cervical spine manual spinal stiffness assessment (in addition reliability of range of motion was reported).	2000-2014 (n=11)†	C°	MSSA	QAREL	<b>Inter-rater reliability</b> Cervical (poor–very good) <b>Intra-rater reliability</b> Cervical (poor–very good)	Data reported; the tests, level/direction, k/ICC, reliability, CI, PA and ROB. Data was graphically presented.	Reliability was collated by specific defined spinal segments. Reliability was reported in k values, CI and PA were reported.
Seffinger et al. (2004) <sup>18</sup> (Systematic review)	To report the reliability of manual spinal stiffness assessment stratified by spinal region.	1976-2001 (n=49)*†	C°, T° and L°	MSSA	Author developed quality criteria with five categories	<b>Inter-rater reliability</b> Cervical (poor–moderate) Thoracic (poor–good) Lumbar (poor–good) <b>Intra-rater reliability</b> Cervical (poor–good)	Data were extracted according to population, number of examiners, examiner discipline, method used to assess and reliability data. Frequencies of associations as per regions, examiners discipline, examination method, asymptomatic/symptomatic etc.	Data pertaining to reporting was extracted. Reliability was reported in k/k(w)/k(m) values, ICC, r and qualitative reporting.

						Thoracic (moderate–good) Lumbar (poor–good)		
Snodgrass et al. (2012) <sup>14</sup> (Systematic review)	To report the reliability of manual spinal stiffness assessment and mechanical assessment stratified by spinal region (in addition to discussing factors associated with spinal stiffness, how stiffness is used in diagnosis, prognosis and treatment decision-making and the effect of spinal manipulation of stiffness).	1987-2012 (n=104)*†	C°, T° and L°	MSSA, M and MA	-	Manual; <b>Inter-rater reliability</b> Cervical (poor–good) Thoracic (poor) Lumbar (fair–good) <b>Intra-rater reliability</b> Thoracic (good)  Mechanical; Test-retest reliability (good)	Data were extracted in a table for the population, method, number of examiners, examiners discipline and main findings.	The purpose of the review was to extract trends and findings from studies. The reliability data wasn't consistently present in the primary studies. Reliability was reported in k values, r and ICC.
Stochkendahl et al. (2006) <sup>43</sup> (Systematic review)	To report the reliability of manual spinal stiffness assessment stratified by spinal region.	1980-2004 (n=31)*†	C°, T° and L°	MSSA	Stochkendahl quality of reproducibility studies	<b>Inter-rater reliability</b> Cervical (poor–good) Thoracic (poor–moderate) Lumbar (poor–good) <b>Intra-rater reliability</b> Cervical (fair–moderate) Thoracic (poor–good) Lumbar (poor–good)	Data were extracted according to population, number of examiners, examiner discipline, method used to assess and reporting used. Meta-analysis was conducted and studies were grouped according to examination method.	Not all studies had reliability data. Reliability data was reported as either k/k(w) values, ICC and PA.
Van Trijffel et al. (2005) <sup>44</sup> (Systematic review)	To report the reliability of manual spinal stiffness assessment method stratified by spinal region.	1982-2004 (n=19)*†	C° and L°	MSSA	Criteria derived from standards for reporting of diagnostic accuracy (STARD) and quality assessment tool for studies of diagnostic accuracy (QUADAS)	<b>Inter-rater reliability</b> Cervical (poor–good) Thoracic (poor) Lumbar (poor–good)	Data were extracted according to population, number of examiners, examiner discipline, method used to assess. reporting used and reliability data.	Author concluded overall interrater reliability poor-substantial. Reliability data was reported as either k/k(w) values, ICC and PA.
Wong et al. (2017) <sup>17</sup> (Narrative review)	To report the reliability of lumbar spine mechanical spinal stiffness assessment (in addition to discussing the reliability of manual and mechanical stiffness assessment, the evidence regarding LBP and spinal stiffness and future clinical research direction).	1990-2016 (n=60)*†	L°	MSSA, M and MA	-	<b>Test-re-test reliability</b> Lumbar (good)	Data were collected about the devices, reliability of the device and the segmental level.	No MSSA data was collected. MSSA data was described in narrative format. No data was collected with regards to examiners discipline or the number of examiners.

\* Symptomatic, † Asymptomatic

Assessment region: cervical (C°), thoracic (T°) and lumbar (L°)

Assessment type: manual spinal stiffness assessment (MSSA), mechanical (M) and mechanically assisted (MA)

Reliability categories: M assessment used k values poor = < 0.2, fair = 0.21-0.4, moderate = 0.41-0.6, good = 0.61-0.80 and very good = 0.91-1.0. M and MA reported ICC values; poor = 0-0.4, fair to good = 0.41-0.75 and good = 0.75-1.

## 2.8 Critical appraisal of included reviews

Table 2.3 reports the critical appraisal of methodological quality for included reviews as determined by two raters (AY, AD) using the AMSTAR-2 checklist. Overall, all included reviews reported search strategy and inclusion criteria, but the reason for excluding citations were infrequently described. Study selection and data extraction methods were variably reported. All but three studies performed risk of bias on included studies<sup>14,17,41</sup>. Only one review performed meta-analysis to synthesis reliability data<sup>43</sup>.

**Table 2.3** AMSTAR-2 checklist

Included review	1. PICO	2. Established protocol prior to review	3. Included study designs	4. Literature search strategy defined	5. Study selection performed in duplicate	6. Data extraction performed in duplicate	7. List of excluded studies	8. Included studies described	9. Risk of bias (RoB) in individual studies	10. Funding disclosed	11. Appropriate methods used for meta-analysis	12. Factor RoB in meta-analysis	13. RoB accounted for in discussion	14. Explanation of heterogeneity	15. Publication bias	16. Conflicts of interests reported
Haneline et al. (2008) <sup>39</sup>	Y	N	Y	Y	Y	N	N	Y	Y	N	NMA	NMA	Y	Y	NMA	N
Hollerwoger (2006) <sup>40</sup>	Y	N	Y	Y	N	N	Y	PY	Y	N	NMA	NMA	Y	PY	NMA	N
Huijbregts (2002) <sup>41</sup>	Y	N	Y	Y	N	N	N	Y	N	N	NMA	NMA	N	Y	NMA	N
Jonsson et al. (2018) <sup>42</sup>	Y	PY	Y	Y	Y	PY	N	Y	Y	PY	NMA	NMA	Y	Y	NMA	Y
Seffinger et al. (2004) <sup>18</sup>	Y	PY	Y	Y	Y	PY	N	Y	Y	PY	NMA	NMA	Y	Y	NMA	N
Snodgrass et al. (2012) <sup>14</sup>	Y	N	Y	Y	Y	N	N	Y	N	Y	NMA	NMA	N	Y	NMA	N
Stochkendahl et al. (2006) <sup>43</sup>	Y	PY	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Van Trijffel et al. (2005) <sup>44</sup>	Y	N	Y	Y	Y	Y	N	Y	Y	N	NMA	NMA	Y	Y	NMA	N
Wong et al. (2017) <sup>17</sup>	Y	N	Y	Y	N	N	N	Y	N	Y	NMA	NMA	N	Y	NMA	Y

Yes (Y), no (N), partial yes (PY), no meta-analysis conducted (NMA), risk of bias (RoB)

## **2.9 Manual spinal stiffness assessment: Inter and intra-examiner reliability**

Eight reviews examined inter-examiner reliability of MSSA. All reviews reported a wide range in reliability across all spinal regions. The cervical spine was examined in all eight reviews of which six reviews reported poor to good reliability<sup>39,41,43,44</sup>, two reviews reported poor to very good reliability<sup>40,42</sup>, and Seffinger et al. (2004) rated inter-examiner reliability of the cervical spine as poor to moderate<sup>18</sup>. The reliability of MSSA for the thoracic spine was examined in five reviews of which three reviews reported that inter-examiner reliability ranged from poor to good<sup>14,39,44</sup>, and two reviews rated thoracic inter-examiner reliability as poor to moderate<sup>41,43</sup>. The inter-examiner reliability of MSSA for the lumbar spine was examined in six review of which five reviews reported poor to good inter-examiner reliability<sup>18,39,41,43,44</sup>, and Snodgrass et al. (2012) reported fair to good inter-examiner reliability for MSSA<sup>14</sup>.

Three reviews found higher levels of inter-rater agreement when stiffness was reported on a dichotomous scale (hypomobile or normal/hypermobile) compared to reporting on an ordinal scale, however the probability of chance agreement is magnified when using a smaller rating scale<sup>14,41,42</sup>. In addition, three reviews reported a positive relationship between primary study quality (based on three different assessment tools), and level of agreement for inter-rater reliability<sup>39,42,43</sup>.

There was a consensus among three reviews that intra-examiner reliability was greater than inter-examiner reliability for MSSA<sup>39,41,43</sup>. Stochkendahl et al. (2006)<sup>43</sup> reported that intra-examiner reliability for MSSA reached clinically acceptable levels when subjects re-identified levels based on the presence of pain ( $\kappa > 0.4$ ).

## **2.10 Mechanical devices: Test-retest reliability**

Mechanical and mechanically assisted device test-retest reliability was explored in two reviews, one review explored all spinal regions<sup>14</sup>, and the other review, lumbar spine only<sup>17</sup>. Both reviews found good test-retest reliability for mechanical methods of spinal stiffness assessment<sup>14,17</sup>. Conclusions regarding the test-retest reliability of mechanical devices were based on reviews that extracted data from 12 primary studies<sup>21,25-30,32,33,52-55</sup>.

## **2.11 Summary of principal findings**

This umbrella review provides a review of reviews that explored the reliability of manual and mechanical methods used to assess spinal stiffness. Overall, there was a consensus among reviews with regard to the wide range (from poor to very good) of intra- and inter-examiner reliability for MSSA across all spinal regions. It remains unclear if the wide range of reliability reported for MSSA is due to the qualitative nature of assessment, issues relating to study quality, or the wide range of study methodologies that include study population, type of test or examiner experience<sup>14,39,41,42,44</sup>.

In contrast to manual assessment, reviews that included mechanical and mechanically assisted devices used for spinal stiffness assessment demonstrated good test-retest reliability<sup>14,17</sup>. The majority of included reviews either used a non-validated quality appraisal tool<sup>18,39,43,44</sup>, or did not appraise study quality<sup>14,17,41</sup> which made comparison of recommendations between reviews difficult.

## **2.12 Summary of methodological issues in the field**

All reviews reported that primary studies had poorly standardised testing procedures, variable reporting outcomes, heterogeneous study populations, and examiners had varied testing procedures<sup>14,17,18,39-44</sup> which limited the ability of reviews to compare studies. Despite the differences in study methodology, one review performed a meta-analysis pooled by spinal region<sup>43</sup>.

Foundational to the assessment of spinal stiffness is the ability to accurately identify surface anatomical landmarks<sup>41</sup>. However, poor inter-examiner reliability for identification of anatomical landmarks was reported in by multiple included reviews<sup>17,18,40-42</sup>. In addition, Hollerwoger (2006)<sup>40</sup> calculated that when examiners were required to agree on a segmental level, the inter-examiner reliability of spinal stiffness assessment was halved. Three reviews agreed that the presence of pain increases the reliability for assessment of spinal stiffness<sup>18,41-43</sup>. An alternate explanation for the increased reliability of landmark detection when pain is present may be the patient's ability to anticipate and recall the location of the pain<sup>42</sup>. Therefore, the reliability of spinal stiffness assessment within a laboratory setting may not

transfer to the clinical setting and depends on the factors such as the ability to accurately identify landmarks and subject symptomatology.

## **2.13 Generalisability**

Five reviews were critical that the study populations were not representative of a clinical population<sup>18,39,41,43,44</sup>, given that a majority of primary studies investigated asymptomatic participants. Moreover, symptomatic populations were only examined in 25 (34%) of primary studies (Appendix C: Primary reliability studies). The reliability of MSSA improved when studies were performed on symptomatic patients, suggesting that MSSA may be more reliable in a clinical setting, compared to a laboratory setting with asymptomatic participants<sup>14,39,41,42</sup>. On balance, any conclusions as to the generalisability of findings across study settings or populations must be made with caution.

## **2.14 Chapter 2 Summary**

This umbrella review provides evidence that there exists a wide range of intra- and inter-rater reliability for MSSA, with the possibility of increased reliability when there is segmental agreement, a symptomatic population and a dichotomous reporting scale. It remains unclear if the wide range of reliability reported for MSSA is due to the qualitative nature of assessment, issues relating to study quality or the wide range of experimental designs. On balance, the results of this review do not support the application of MSSA for assessment of spinal stiffness in the research setting, and limited support for use of MSSA in the clinical setting. In contrast, mechanical and mechanically assisted devices offer a reliable method to quantify spinal stiffness.

## 3. Methods: Bench-top performance of the VerteTrack system

### 3.1 Context

The first stage in testing the performance of a novel measurement device is to conduct a series of accuracy experiments under controlled conditions. An accuracy experiment determines the closeness of agreement between a test result and an accepted reference value<sup>38</sup>. Accuracy can be described by a combination of the precision (quantity of random error) and bias (quantity of systematic error) of a measurement system<sup>36</sup>. To measure the precision of a system, the variance of repeated independent measurements under controlled conditions are used<sup>36-38</sup>. Correspondingly, to determine the measurement bias of a system, the mean of repeated measurements are tested for both level of agreement and linearity against a range of reference values.

The aim of this study is to measure the bench-top performance of the VerteTrack system, a novel measurement device for quantifying spinal stiffness.

The hypotheses are:

1. The bench-top performance of the VerteTrack system exhibits appropriate levels of accuracy for future validation testing using human subjects.
2. The bench-top performance of the VerteTrack system has equivalent performance during both static and dynamic modes of operation.

Chapters 3 and 4 describe methods and results respectively of three discrete experiments to address this aim.

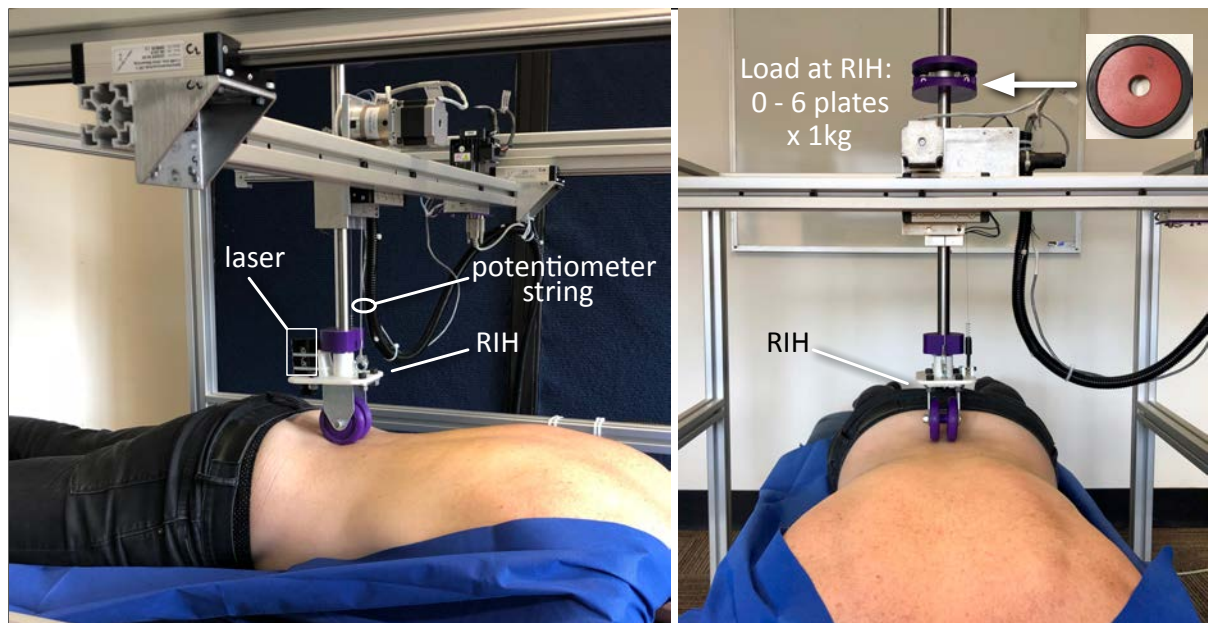
### 3.2 Study design, setting and equipment

The study design is a laboratory-based accuracy experiment<sup>36</sup>. Experiments were designed and conducted in accordance with the International Organization for Standardization (ISO 5725-1) for the accuracy of measurement methods and results<sup>36</sup>. All experiments were

conducted in February 2018 in a dedicated laboratory space located in the Department of Chiropractic, Macquarie University, Sydney.

### 3.2.1 The VerteTrack

The VerteTrack consists of an aluminium gantry (Width 1080 mm × Height 1090 mm × Length 1510 mm) that translates a rolling indentation head along three axes: X-axis (longitudinal, superior-inferior), Y-axis (transverse, left-right) and Z-axis (vertical, posterior-anterior). Stepper motors execute movement about the X, Y and Z-axis (www.stepperonline.com, China, resolution = 0.007mm) (Figure 1.4). Vertical displacement in spinal tissues along the Z-axis is measured by a string potentiometer (TE Connectivity, USA, resolution = 0.020 mm). A vertically-oriented laser mounted on the rolling indenter head (RIH) assists the operator in the alignment of the device to the targeted location/s for indentation (GLX Laser Site, Barska). During spinal stiffness assessment, the VerteTrack applies discrete incremental loads which comprise a series of weightlifting plates (“plates”) with a nominal mass of 1kg. The loads are applied to the Z-axis via the RIH to participants spinal tissues. There are seven possible loads that can be applied to an individual’s lumbar spine (RIH +  $k$  plates;  $k=0, 6$ ). Plates were serialised (labelled one through six) and were added in the same order for each indentation cycle (Figure 3.1).

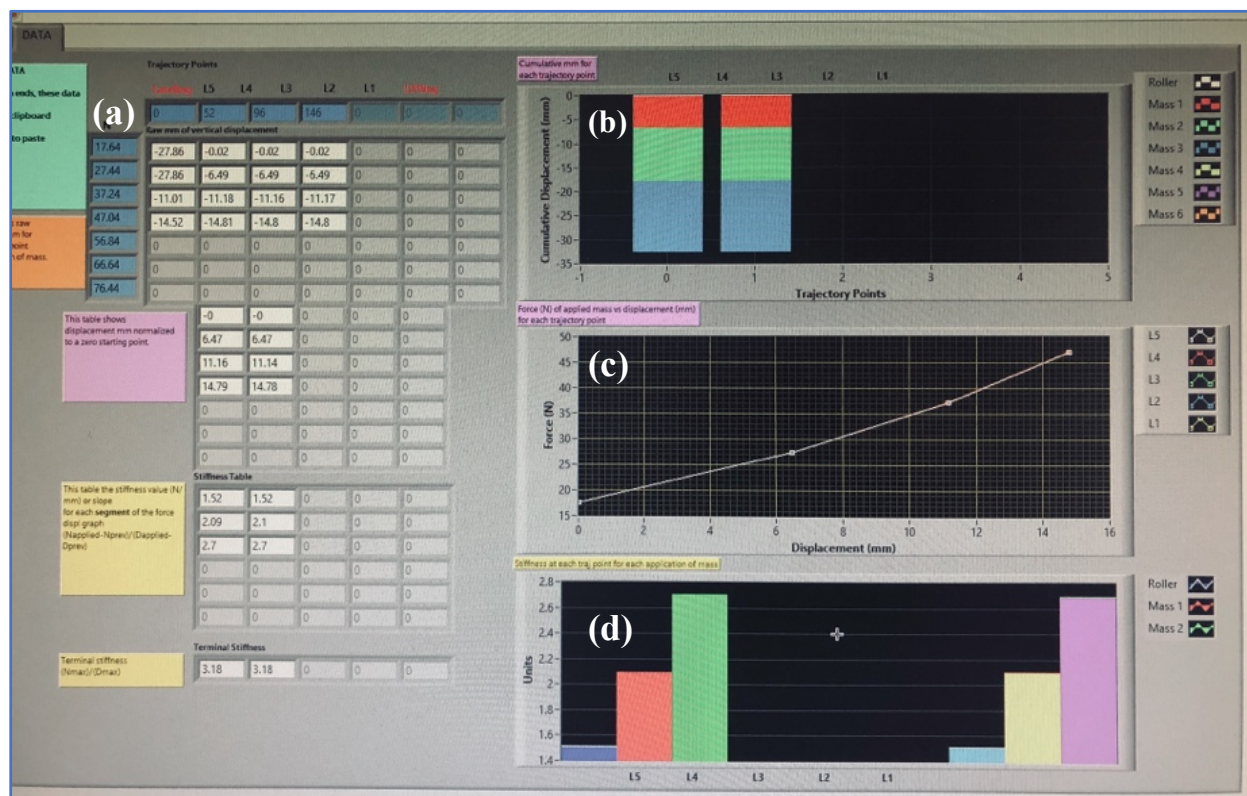


**Figure 3.1** A labelled image of the VerteTrack during indentation.  
RIH – Rolling indenter head



### 3.2.2 VerteTrack custom software

Data is acquired from the X, Y and Z-axis through customised software on Labview (National Instruments, USA). The software receives and stores information via encoders embedded in the X and Y-axis. Stepper motors enable the indenter head to follow a pre-mapped trajectory about the X and Y-axis. Labview retrieves vertical displacement data from the Z-axis via the string potentiometer. The main information screen displays a real-time capture of stiffness data (Figure 3.2). Raw data is tabulated and can be presented graphically (i.e. force-displacement curve) through Labview. The typical force-displacement profile shows a smooth curvilinear line graph showing a positive relationship between load and displacement. Albeit, when indentation is performed on participants the initial loads (smaller loads) cause minimal displacement in spinal tissues, referred to as a ‘toe region’<sup>55,56</sup>.



**Figure 3.2** Data analysis software

- (a) Raw tabulated load (N) and displacement (mm) data for different locations
- (b) Cumulative displacement (mm) for different locations
- (c) Force (N) versus displacement (mm) curve displayed for one location (F-D curve)
- (d) Stiffness (N/mm) histogram for each location

### 3.2.3 Indentation procedure

The VerteTrack can perform two modes of indentation testing: static and dynamic. Static indentation assesses a single spinal level and requires the operator to manually position the RIH directly above the target tissue with the assistance of the RIH mounted laser.

Incremental posterior to anterior loads are applied to participants spinal tissues which result in deformity and Z-axis displacement. The load and displacement variables are exported into Labview to produce force-displacement curves and terminal stiffness values. Static indentation aligns with previous mechanical and mechanically assisted device (single level indentation), which only allow posterior to anterior translation of the device. This process is designed to mimic manual spinal stiffness assessment.

Dynamic indentation requires the operator to first plan the trajectory of RIH movement in X, Y and Z-axes. This is achieved by first placing the VerteTrack in a “training” mode, whereby the operator manually traces anatomical surface landmarks (spinous processes) of the participants using the RIH mounted laser (GLX Laser Site, Barska). When the laser aligns with the pre-determined locations, the position is recorded and the indenter head trajectory is calculated. This pre-mapped trajectory is termed a trace. For dynamic indentation, once the trace is determined and the initial load (RIH only) is applied to the VerteTrack indentation can begin. The indenter head travels down to the first location on the participant or medium termed ‘landing point’ then follows the trace to the final location which is termed the ‘lift-off’ point. This process is then repeated for the six subsequent loads. During bench-top performance testing, simulated surface landmarks were drawn on the test medium, then used for multiple cycles of both static and dynamic indentation. Table 3.1 provides operational definitions used for indentation procedures.

**Table 3.1** Operational definitions used for indentation procedures

Term	Operational definition
Rolling Indenter Head (RIH)	The component of the VerteTrack in contact with the target tissue
Indentation	The action of the indenter head landing, loading, then rising from the target tissue (translation along Z-axis) during: <ul style="list-style-type: none"> <li>○ <i>Static indentation</i>: single spinal level per indentation (indentation at fixed X and Y-axes)</li> <li>○ <i>Dynamic indentation</i>: multiple spinal levels per indentation (indentation with translation along <math>X \pm Y</math>-axes)</li> </ul>
Trace	A pre-mapped dynamic indentation trajectory
One trial	One indentation at a single load
One cycle	A sequence of seven trials at increasing load
Terminal stiffness	Calculated as the ratio between maximum load and maximum displacement

### 3.3 Experiment one: Precision and bias of the VerteTrack for the applied load

Experiment one evaluated the accuracy (precision and bias) of loads applied through the RIH, compared to reference values.

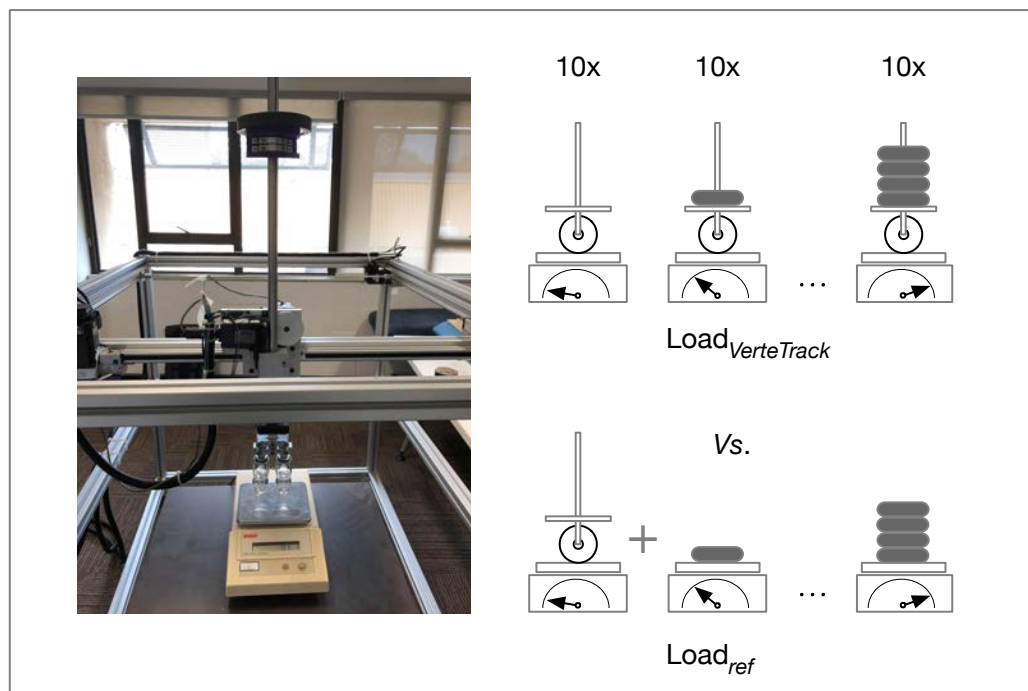
#### 3.3.1 Precision of load

A digital scale (OHAUS, model TS4KD: Resolution 0.1g, accuracy  $\pm 0.07\text{g}$ ) was placed underneath VerteTrack (Figure 3.3). The load applied at the VerteTrack indenter head (RIH) (without plates) was measured by lowering then raising the RIH ten times ( $\text{Load}_{\text{VerteTrack}}$ ). One plate was added to the RIH, then repeated as above up to a total of 4 plates ( $\text{RIH} + k$  plates;  $k=0, 4$ ). Each plate was labelled with a serial number. Loads were converted to Newtons (N) using mass (kg)  $\times$  gravity ( $9.81\text{m/s}^2$ ). Loads were estimated with 95% confidence and standard deviation (SD) for each of the 5 discrete loads ( $\text{Load}_{\text{VerteTrack}}$ ). The coefficient of variation (CV) was calculated for each load to estimate the dispersion (Equation 2).

$$CV = \left( \frac{SD}{\text{mean}(\text{Load}_{\text{VerteTrack}})} \right) \times 100 \quad \text{Equation 3.2}$$

### 3.3.2 Bias of load

Each load ( $\text{Load}_{\text{VerteTrack}}$ ) from 3.3.1 was compared to a corresponding reference value. Reference loads ( $\text{Load}_{\text{ref}}$ ) were calculated by the addition of successive plates placed directly upon the digital scale (i.e. not through the VerteTrack RIH) plus the load measured through the RIH alone (from 3.3.1). Each reference load ( $k$  plates;  $k=1, 4$ ) was measured 10 times. The mean systematic bias for  $\text{Load}_{\text{VerteTrack}}$  vs.  $\text{Load}_{\text{ref}}$  was measured using the Bland-Altman limits of agreement to capture 95% of the differences between the two measurements and visualised using the Bland-Altman plot<sup>57</sup>. In addition, Lin's Concordance Correlation Coefficient ( $\text{LinCCC}(R_c)$ ) was reported which is an omnibus statistic to test both agreement and linearity for the four discrete loads  $\text{Load}_{\text{VerteTrack}}$  vs.  $\text{Load}_{\text{ref}}$ <sup>58</sup>. The strength of agreement was graded as "almost perfect" ( $R_c > 0.99$ ), "substantial" ( $R_c > 0.95-0.99$ ), "moderate" ( $R_c > 0.90-0.95$ ), or "poor" ( $R_c < 0.90$ )<sup>59</sup>. Figure 3.3 illustrates the methodology for experiment one.



**Figure 3.3** Experiment one methodology: load applied by the VerteTrack ( $\text{Load}_{\text{VerteTrack}}$ ) vs. reference load ( $\text{Load}_{\text{ref}}$ )

## 3.4 Experiment two: Precision and bias of the VerteTrack indenter head displacement

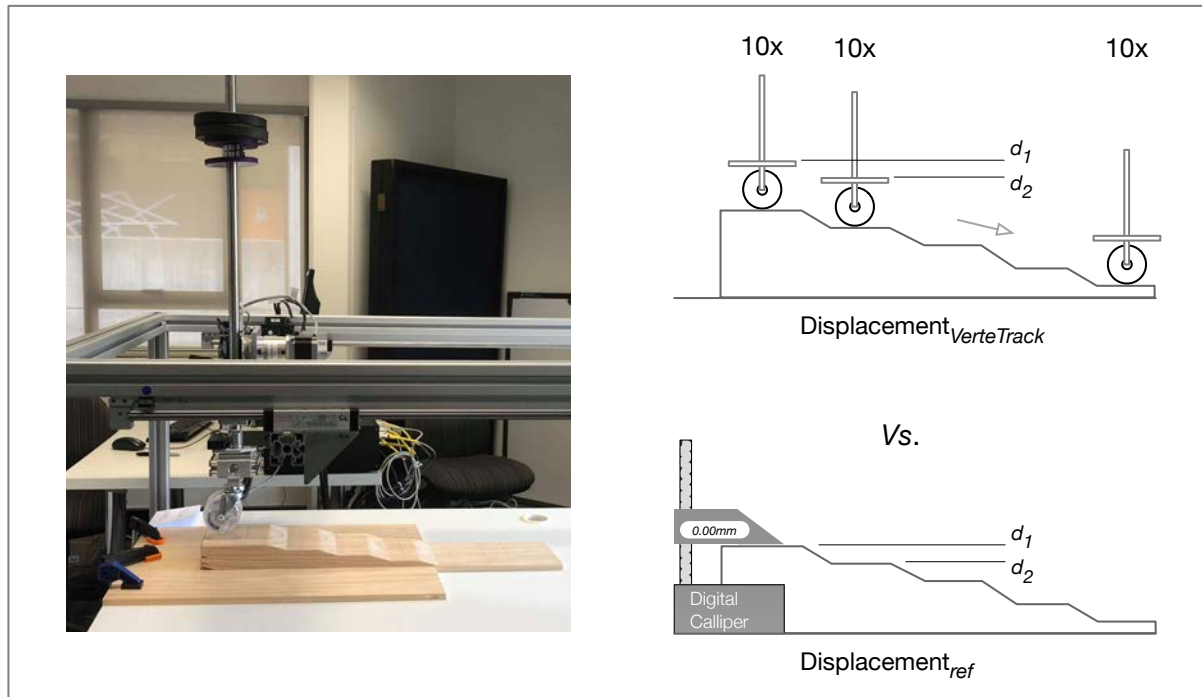
Experiment two evaluated the accuracy (precision and bias) of RIH displacement as recorded by the VerteTrack, compared to reference values. The magnitude of the RIH displacement due to an applied load determines stiffness of the test medium.

### 3.4.1 Precision of VerteTrack RIH displacement

A string potentiometer (TE Connectivity, USA, Resolution: 0.020mm, accuracy  $\pm 0.010$ mm) mounted on the Z-axis recorded displacement at the RIH at each of 6 discrete levels on a custom engineered wooden wedge (Figure 3.4). The magnitude of displacement was measured relative to table-top upon which the wedge was securely clamped. Displacement was measured ten times at each wedge level using VerteTrack ( $\text{Displacement}_{\text{VerteTrack}}$ ) and comprised the landing point ( $d_0$ ) and the highest level ( $d_1$ ), followed by one location for each subsequent lower level ( $d_2 - d_5$ ). Displacement from the table-top was estimated with 95% confidence, standard deviation (SD) and coefficient of variation (CV).

### 3.4.2 Bias of VerteTrack RIH displacement

Each displacement measurement ( $\text{Displacement}_{\text{VerteTrack}}$ ) in 3.4.1 was compared to a reference value obtained by a digital calliper (Wixey, WR200: Resolution = 0.05mm, accuracy  $\pm 0.025$ mm). Each reference level was measured 10 times. The mean systematic bias for  $\text{Displacement}_{\text{VerteTrack}}$  and  $\text{Displacement}_{\text{ref}}$  was measured using the Bland-Altman limits of agreement to capture 95% of the differences between the two measurements, and visualised using the Bland-Altman plot. In addition, Lin's Concordance Correlation Coefficient ( $\text{LinCCC}(R_c)$ ) was reported to test both agreement and linearity for the six discrete displacements across two devices (VerteTrack vs. digital calliper as reference). The strength of agreement from LinCCC was rated as per criteria used in experiment one<sup>59</sup>.



**Figure 3.4** Experiment two methodology: displacement measured by the VerteTrack ( $\text{Displacement}_{\text{VerteTrack}}$ ) vs. reference displacement ( $\text{Displacement}_{\text{ref}}$ )

### 3.5 Experiment three: Performance of the VerteTrack system during both static and dynamic modes of operation

Experiment three was a method comparison experiment to evaluate the performance of VerteTrack for measurement of stiffness during dynamic and static modes of operation. Dynamic stiffness profiles were compared to static (reference) stiffness profiles. Terminal stiffness values (the ratio of the maximum load to the maximum displacement<sup>31</sup>) were used for analysis in experiment three.

#### 3.5.1 Methods of indentation and medium characteristics

The stiffness of a deformable test medium (AIREX balance beam, Switzerland) was measured using both static and dynamic modes of operation to simulate measurement at a single vertebral level and across multiple vertebral levels respectively. The test medium was chosen to emulate the physiological stiffness encountered for the in vivo adult lumbar spine (range: 2 – 10 N/mm)<sup>26,31,60</sup>. Additionally, the AIREX balance beam is designed to exhibit minimal indentation force deflection (IFD) fatigue at loads in excess of 100kg over a small contact area (i.e. it is designed to be walked upon without permanent deformity), meaning that loads applied by VerteTrack will be unlikely to cause IFD fatigue in the AIREX balance

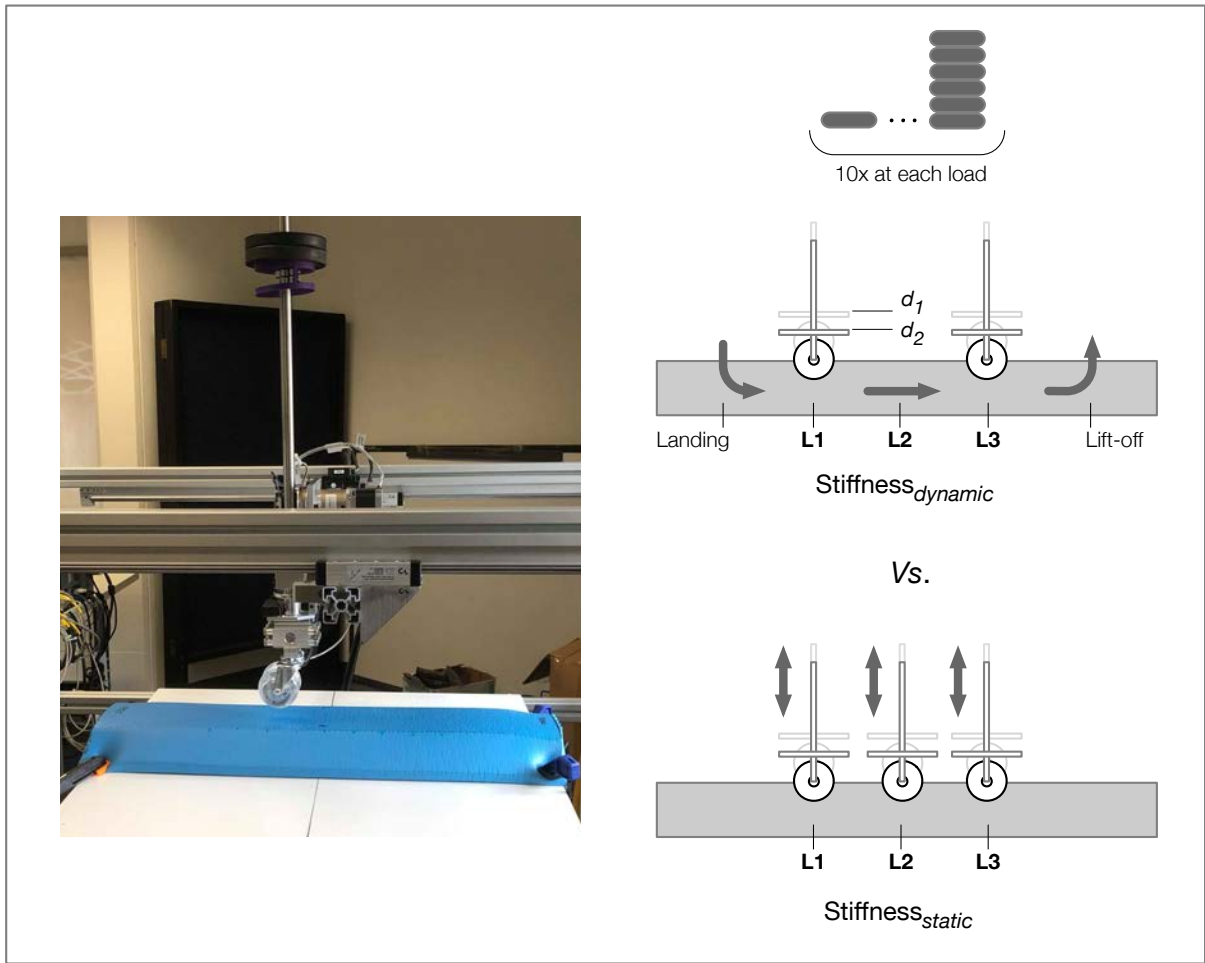
beam. Five equidistant locations (5cm apart) were marked on the foam medium along a straight line (landing, L1, L2, L3 and lift-off) for stiffness assessment (Figure 3.5).

#### Static indentation

Stiffness was measured using static indentation ( $Stiffness_{static}$ ) at three discrete locations (L1, L2 and L3) on the medium. Incremental loads (plates) were added to the RIH in a predefined sequence ( $RIH + k$ ;  $k=1, 6$ ), with each plate serialised for re-identification. Between each trial, 90 seconds elapsed to allow for indentation force deflection to return to zero. Between each cycle (six trials of increasing load), an additional 5 minutes elapsed to allow any residual deformation to resolve after the maximum load was applied to the medium. A total of ten cycles were performed. Terminal stiffness from each trial was extracted from the raw  $Stiffness_{static}$  data to represent reference stiffness values for locations L1, L2 and L3. (Figure 3.5)

#### Dynamic indentation

In a similar method to static indentation, the same locations were measured using dynamic indentation ( $Stiffness_{dynamic}$ ). Incremental loads (plates) were added to the rolling indenter head in a predefined sequence ( $RIH + k$ ;  $k=1, 6$ ), with each plate serialised for re-identification. Between each trial (represented by landing to lift-off for a single load), 90 seconds elapsed to allow for any IFD to return to zero. Between each cycle (six trials of increasing load), an additional 5 minutes elapsed to allow any residual deformation to resolve after the maximum load was applied to the medium. A total of 10 cycles were performed. Terminal stiffness from each trial was extracted from the raw stiffness data to represent  $Stiffness_{dynamic}$  at locations L1, L2 and L3 (Figure 3.5).



**Figure 3.5** Experiment three methodology: stiffness measurements using dynamic indentation ( $Stiffness_{dynamic}$ ) vs. static indentation ( $Stiffness_{static}$ )

### 3.5.2 Precision of stiffness

Stiffness values at the three discrete locations (L1, L2 and L3) for  $Stiffness_{static}$  and  $Stiffness_{dynamic}$  were used to generate standard deviation (SD) and coefficient of variation (CV), which were estimated with 95% confidence.

### 3.5.3 Bias of stiffness

Each stiffness measurement for  $Stiffness_{dynamic}$  obtained in 3.5.2 was compared to  $Stiffness_{static}$ . The mean systematic bias for  $Stiffness_{dynamic}$  and  $Stiffness_{static}$  was compared using the Bland-Altman limits of agreement to capture 95% of the differences between the two measurements, and visualised using the Bland-Altman plot. To further assist with the interpretation of bias, a plot of raw stiffness data and force-displacement curves were generated.



### **3.6 Project management**

The VerteTrack is housed in the Department of Chiropractic, Faculty of science and engineering at Macquarie University. Materials required for this project were funded by the Macquarie University Masters of Research allocation and was used to purchase; a digital calliper (Wixey, WR200: Resolution = 0.05mm, accuracy  $\pm 0.025\text{mm}$ ), the displacement medium (supplied by AD) and the stiffness medium (AIREX, Balance beam). Data from experiments 1, 2 and 3 were collected by AY with support from GK in February 2018. Raw data was managed by AY on three separate spreadsheets for experiment one- load, experiment two- displacement, and experiment three- terminal stiffness values for dynamic and static. Statistical analysis of precision was generated using Microsoft Excel (2016) by AY. Statistical analysis used to determine bias (LinCCC and Bland-Altman plot), F-D curves and stiffness profiles were generated using IBM SPSS Statistics for Mac, version 25 by AY.

## 4. Results

### 4.1 Experiment one: Precision and bias of the VerteTrack for the applied load

#### 4.1.1 Precision of load

The magnitude of the estimate for five discrete loads measured at the rolling indenter head (RIH) ( $\text{Load}_{\text{VerteTrack}}$ ) ranged from 17.247N (95%CI 17.167 to 17.327, RIH only) to 61.263N (95%CI: 61.211 to 61.316, RIH + 4 plates) (Table 4.1). As each plate was added to the RIH, the  $\text{Load}_{\text{VerteTrack}}$  increased by approximately 9.81 Newtons (nominal weight of plate =1kg). The coefficient of variation (CV) ranged from 0.07% to 0.58% depending upon load (Table 4.1).

**Table 4.1** Precision of VerteTrack applied load

Indenter head loading	$\text{Load}_{\text{VerteTrack}}^*$ (N)	95%CI	SD	CV
RIH only	17.247	(17.167 to 17.327)	0.066	0.38%
RIH + 1 plate	28.448	(28.368 to 28.528)	0.066	0.23%
RIH + 2 plates	39.274	(39.235 to 39.314)	0.032	0.08%
RIH + 3 plates	50.079	(49.726 to 50.432)	0.291	0.58%
RIH + 4 plates	61.263	(61.211 to 61.316)	0.043	0.07%

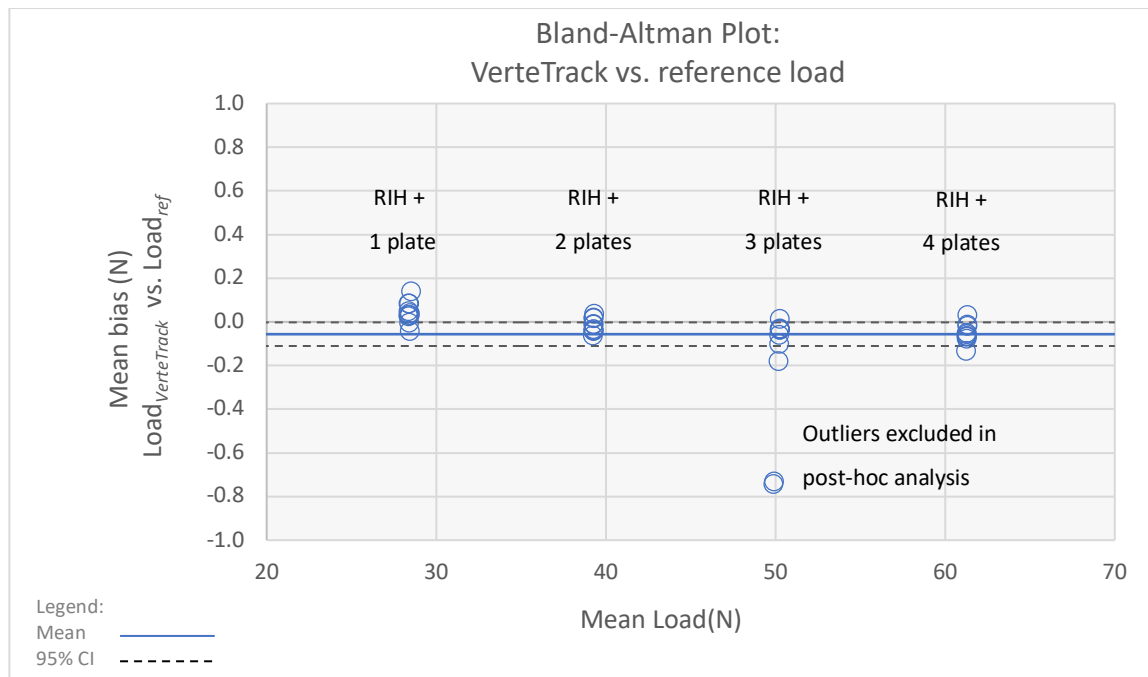
RIH – Rolling indenter head

\*Average of 10 measurements at each load. All loads measured with digital scale (OHAUS, model TS4KD: Resolution =0.1g, accuracy  $\pm 0.07$ g. Equivalent to resolution =0.001N, accuracy  $\pm 0.0007$ N).

#### 4.1.2 Bias of load

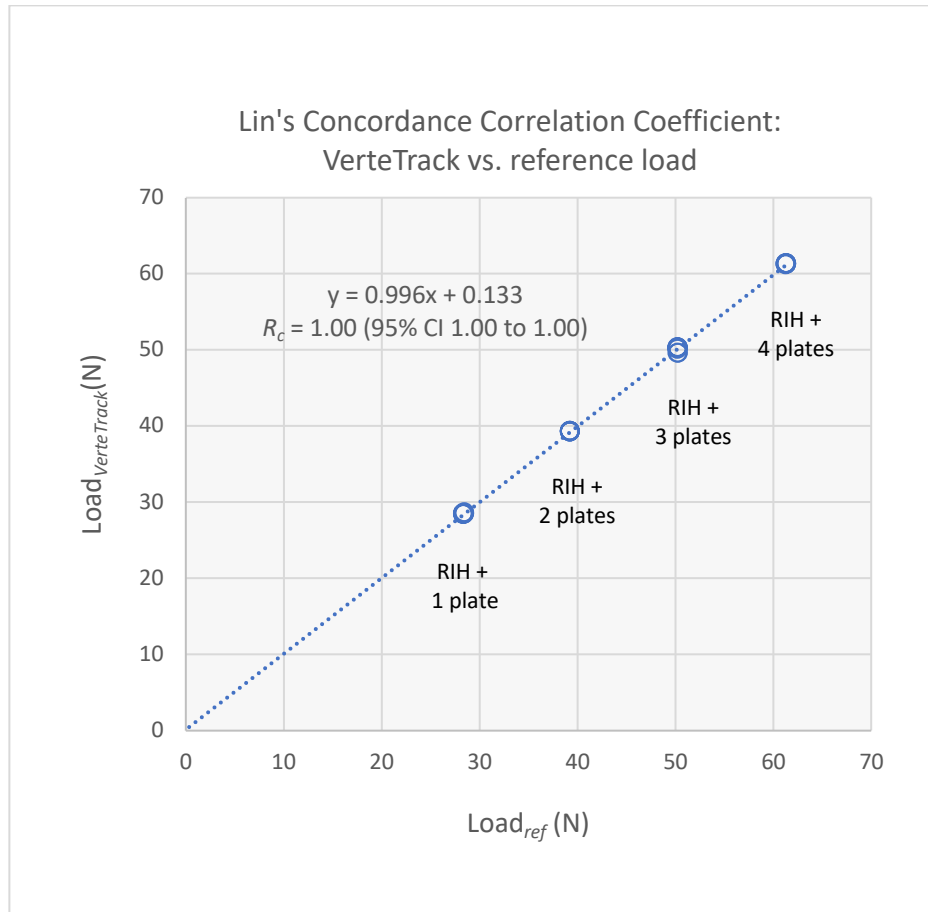
The calculated reference loads ( $\text{Load}_{\text{ref}}$ ) ranged from 28.406N (95%CI 28.348 to 28.464, 1 plate + RIH) to 61.317N (95%CI 61.315 to 61.320, 4 plates + RIH). The Bland-Altman limits of agreement demonstrated a statistically significant ( $p < 0.05$ ) systematic mean bias for the VerteTrack load ( $\text{Load}_{\text{VerteTrack}}$ ), compared to the reference load ( $\text{Load}_{\text{ref}}$ ) of -0.050N (95%CI -0.094 to -0.007;  $t(49) = -2.308$ ,  $p = 0.025$ )(Figure 4.1). There were two outliers observed (RIH+3 plates), which may have been due to transcription error when manually recording load from the scale (Figure 4.1). A post-hoc analysis with outliers excluded demonstrated statistically significant ( $p < 0.05$ ) bias for loads delivered by the VerteTrack

compared to the reference load, but with narrower limits of agreement and smaller magnitude of systematic mean bias ( $-0.021\text{N}$ , 95%CI  $-0.041$  to  $-0.001$ ;  $t(37) = -2.084$ ,  $p = 0.044$ ).



**Figure 4.1** Bland-Altman plot to demonstrate a statistically significant ( $p < 0.05$ ) bias for loads delivered by the VerteTrack compared to the calibration sample ( $-0.050\text{N}$ , 95%CI  $-0.094$  to  $-0.007$ ,  $p = 0.025$ ). Open circles (40 data points) represent the magnitude of bias ( $N = \text{Load}_{\text{ref}} - \text{Load}_{\text{VerteTrack}}$ ).

Lin's Concordance Correlation Coefficient ( $\text{LinCCC}(R_c)$ ) for  $\text{Load}_{\text{VerteTrack}}$  vs.  $\text{Load}_{\text{ref}}$  was 1.00 (95% CI 1.00 to 1.00), which was graded as almost perfect agreement ( $R_c > 0.99$ ) between VerteTrack and reference, across a representative range of loads delivered by the VerteTrack. The line of best fit exhibited high linearity and a (theoretical) y-intercept of  $+0.133\text{N}$  (Figure 4.2).



**Figure 4.2** Lin's Concordance Correlation Coefficient for VerteTrack load vs. the reference sample to demonstrate almost perfect agreement ( $R_c = 1.00$ , 95% CI 1.00 to 1.00). Open circles (40 data points) represent co-ordinates ( $\text{Load}_{ref}$ ,  $\text{Load}_{VerteTrack}$ ) at loads (RIH +  $k$  plates;  $k=1, 4$ )

## 4.2 Experiment two: Precision and bias of the VerteTrack indenter head displacement

### 4.2.1 Precision of RIH displacement

The estimates for six discrete levels were measured at the indenter head ( $\text{Displacement}_{\text{VerteTrack}}$ ) ranged from 60.03mm (95%CI 60.01 to 60.05, highest level) to 12.08mm (95%CI 12.00 to 12.16, lowest level). The coefficient of variation (CV) ranged from 0.01% to 0.32% depending upon the level of the wedge (Table 4.2).

**Table 4.2** Precision of the VerteTrack RIH displacement

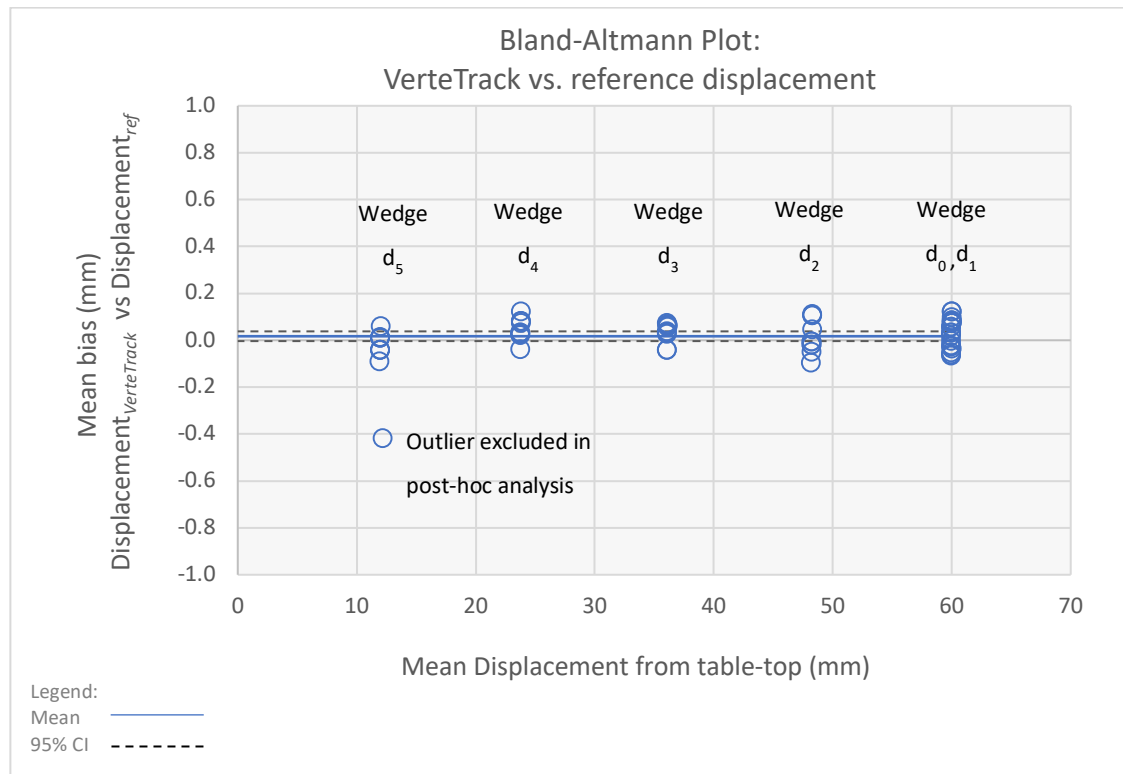
Wedge level	RIH displacement relative to table-top* (mm)	95%CI	SD	CV
d <sub>0</sub> (landing point)	60.03	(60.01-60.05)	0.04	0.01%
d <sub>1</sub> (highest level)	60.02	(60.02-60.02)	0.01	0.00%
d <sub>2</sub>	48.30	(48.29-48.31)	0.01	0.00%
d <sub>3</sub>	36.13	(36.12-36.13)	0.01	0.01%
d <sub>4</sub>	23.82	(23.82-23.83)	0.01	0.01%
d <sub>5</sub> (lowest level)	12.08	(12.00-12.16)	0.14	0.32%

RIH – Rolling indenter head

\*Average of 10 measurements at each displacement. All displacements were measured by the string potentiometer relative to the table-top (TE Connectivity, USA, Resolution: 0.020mm, accuracy  $\pm 0.010$ mm)

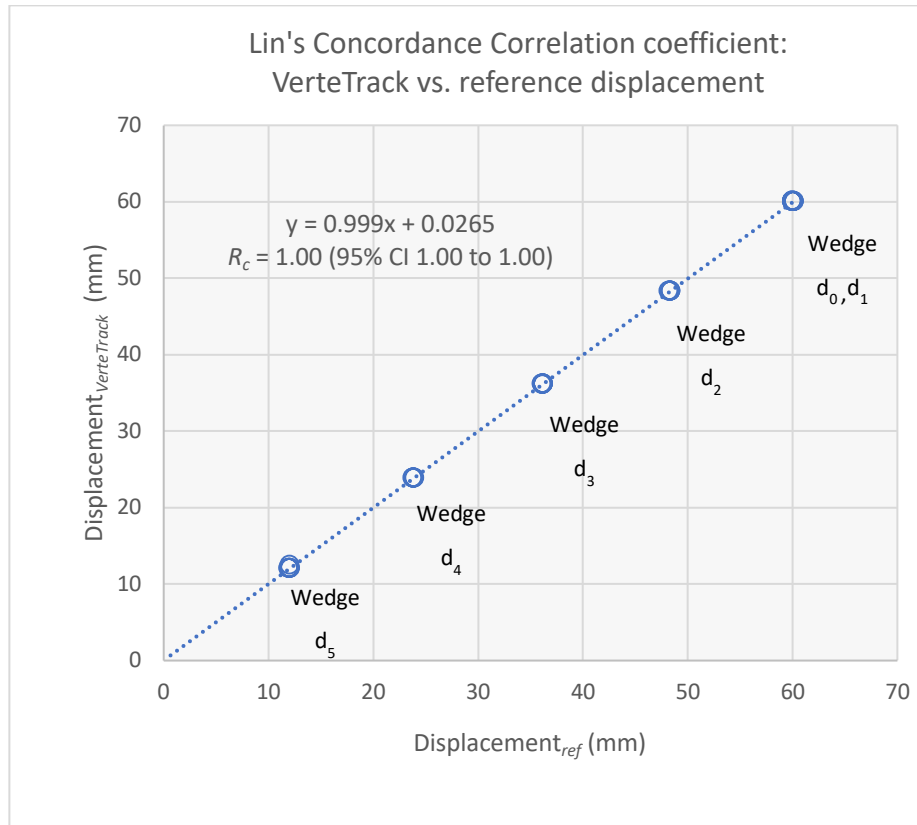
### 4.2.2 Bias of RIH displacement

The reference displacement ( $\text{Displacement}_{\text{ref}}$ ) was measured by the digital calliper (resolution = 0.050mm, accuracy  $\pm 0.025$ mm) ranged from 60.08mm (95%CI 60.02 to 60.13, location d<sub>1</sub>), to 12.03mm (95%CI 11.98 to 12.08, location d<sub>5</sub>). The Bland-Altman limits of agreement demonstrated non-significant ( $p > 0.05$ ) systematic bias for the VerteTrack displacement ( $\text{Displacement}_{\text{VerteTrack}}$ ), compared to the reference displacement ( $\text{Displacement}_{\text{ref}}$ ) (mean difference = 0.02mm, 95%CI 0.00 to 0.04;  $t(59) = 1.66$ ,  $p = 0.102$ ) (Figure 4.3). There was one outlier observed at the lowest wedge level (d<sub>5</sub>), which may have been due to transcription error when manually recording displacement from the digital calliper (Figure 4.3). A post-hoc analysis with outlier excluded demonstrated statistically significant ( $p < 0.05$ ) bias for displacements measured by the VerteTrack compared to the digital calliper sample (0.02mm, 95%CI 0.01 to 0.4;  $t(58) = 3.334$ ,  $p = 0.001$ ).



**Figure 4.3** Bland-Altman plot to demonstrate statistically non-significant bias ( $p > 0.05$ ) for displacement as measured by the VerteTrack compared to the digital calliper (+0.02mm, 95%CI 0.00 to 0.04,  $p = 0.102$ ). Open circles (60 data points) represent the magnitude of bias (mm) =  $\text{Displacement}_{ref} - \text{Displacement}_{VerteTrack}$

Lin's Concordance Correlation Coefficient ( $\text{LinCCC}(R_c)$ ) for  $\text{Displacement}_{VerteTrack}$  vs.  $\text{Displacement}_{ref}$  was 1.00 (95% CI 1.00 to 1.00), which was graded as almost perfect agreement ( $R_c > 0.99$ ) between VerteTrack and reference across the full range of measured displacements (Figure 4.4). The line of best fit exhibited high linearity and a (theoretical) y-intercept of +0.03mm (Figure 4.4). These results must be interpreted within the context of the resolution of the reference measuring device (digital calliper: resolution = 0.050mm, accuracy  $\pm 0.025$ mm).



**Figure 4.4** Lin's Concordance Correlation Coefficient for VerteTrack displacement vs. the digital calliper demonstrated an almost perfect agreement ( $R_c = 1.00$ , 95% CI 1.00 to 1.00). Open circles (60 data points) represent co-ordinates ( $\text{Displacement}_{\text{ref}}$ ,  $\text{Displacement}_{\text{VerteTrack}}$ ) for each wedge level ( $d_0$ – $d_5$ )

## 4.3 Experiment three: Performance of the VerteTrack system during both static and dynamic modes of operation

### 4.3.1 Precision of stiffness

To determine the precision of stiffness measured by the VerteTrack ( $Stiffness_{dynamic}$  and  $Stiffness_{static}$ ) at three locations on the AIREX balance beam the coefficient of variation was calculated for each load and shown in Table 4.3. The coefficient of variation at each load for  $Stiffness_{static}$  ranged from 2.0% to 2.3% and  $Stiffness_{dynamic}$  ranged from 1.4% to 3.2%.

**Table 4.3** Coefficient of variation for  $Stiffness_{dynamic}$  and  $Stiffness_{static}$

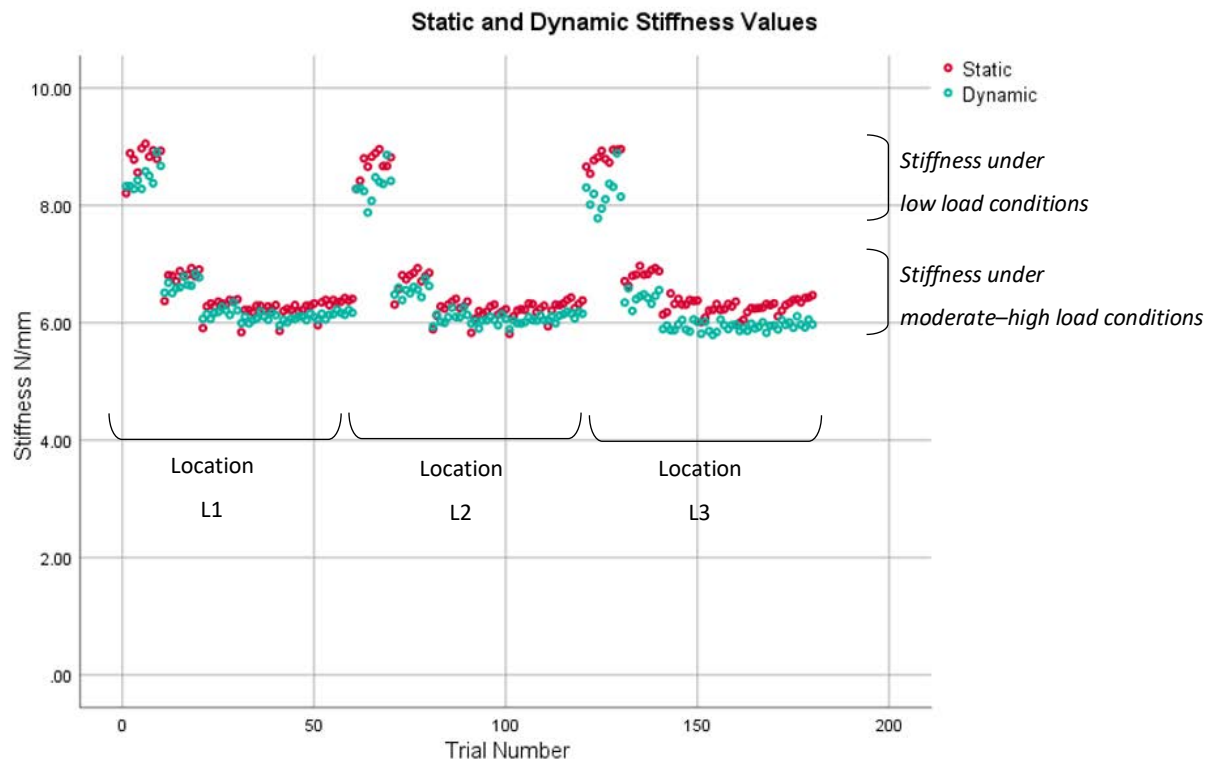
VerteTrack mode	Coefficient of variation at each load						Mean CV
	RIH+1	RIH+2	RIH+3	RIH+4	RIH+5	RIH+6	
Static	2.3%	2.2%	2.1%	2.0%	2.1%	2.0%	<b>2.1%</b>
Dynamic	3.2%	2.2%	2.2%	1.7%	1.4%	1.6%	<b>2.1%</b>

RIH – Rolling indenter head

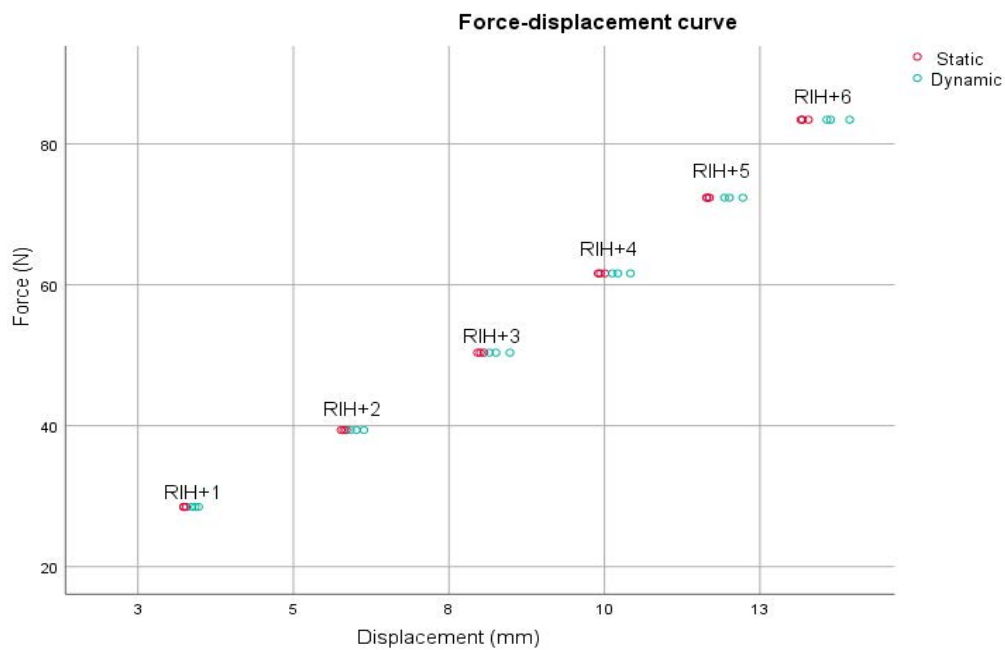
### 4.3.2 Agreement of VerteTrack Static and Dynamic stiffness measurements

To evaluate agreement, dynamic indentation ( $Stiffness_{dynamic}$ ) and static indentation ( $Stiffness_{static}$ ) were compared at three discrete locations (L1, L2 and L3) on the AIREX balance beam. A plot of the raw data for all 3 locations is shown in Figure 4.5. Stiffness values recorded under low load conditions (RIH+1 and RIH+2) exhibited consistently greater magnitude. In addition, dynamic ( $Stiffness_{dynamic}$ ) and static ( $Stiffness_{static}$ ) were compared using a force-displacement curve (Figure 4.6) for locations (L1, L2 and L3). Dynamic indentation consistently showed greater displacement in the medium compared to static indentation.



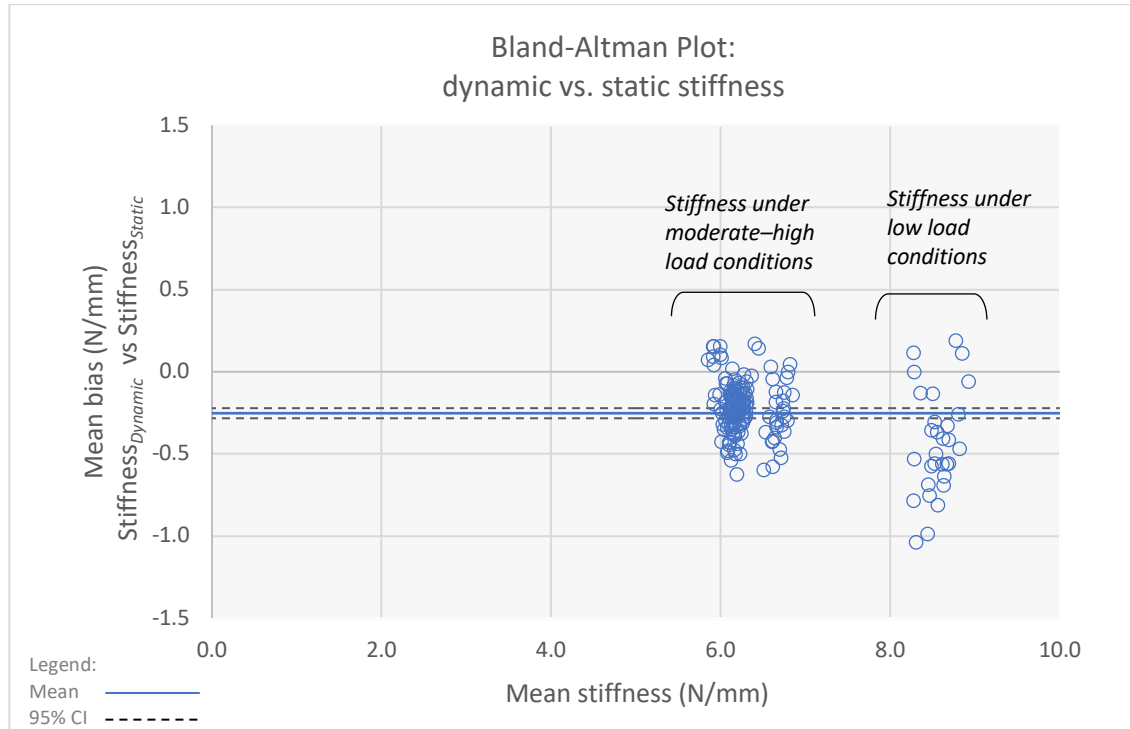


**Figure 4.5** Dynamic and static stiffness data across three locations (L1, L2 and L3) on the AIREX balance beam. L1: trial 0 – 60, L2: trial 61 – 120, L3: trial 121 – 180



**Figure 4.6** Force-displacement curve for  $Stiffness_{dynamic}$  and  $Stiffness_{static}$ . Mean displacement at loads (RIH + k plates; k=1, 6) for all three locations (L1, L2 and L3) (36 data points)

The Bland-Altman plot demonstrated a statistically significant ( $p < 0.05$ ) negative systematic bias for stiffness calculated using dynamic indentation, compared to stiffness calculated using static indentation of  $-0.25 \text{ N/mm}$  (95%CI  $-0.28$  to  $-0.22$ ;  $t(179) = 16.2$ ,  $p < 0.001$ ) (Figure 4.7).



**Figure 4.7** Bland-Altman plot demonstrating the negative bias for dynamic vs. static stiffness. Bland-Altman plot to demonstrate a statistically significant ( $p < 0.05$ ) bias for stiffness calculated using dynamic indentation, compared to stiffness calculated using static indentation of  $-0.25 \text{ N/mm}$  (95%CI  $-0.28$  to  $-0.22$ ,  $p < 0.001$ ). Open circles (180 data points) represent the magnitude of bias ( $N$ ) =  $\text{Stiffness}_{\text{dynamic}} - \text{Stiffness}_{\text{static}}$

# 5. Discussion

## 5.1 Overview of main findings

This thesis provides new insights and original contributions to the field of devices used to assess spinal stiffness. This thesis establishes the reliability of existing assessment procedures (Chapter 2) and investigates the bench-top performance of a new device, the VerteTrack, for the assessment of spinal stiffness (Chapter 3 and 4).

### 5.1.1 Reliability of existing spinal stiffness assessment methods

The first aim of this thesis was to determine the reliability of current methods used for spinal stiffness assessment (Chapter 2). The umbrella literature review examined the intra- and inter-rater reliability of manual spinal stiffness assessment (MSSA), and test-retest reliability of mechanical and mechanically assisted devices. There was uniformity among included reviews with regards to the wide range (from poor to very good) of intra- and inter-examiner reliability for MSSA across all spinal regions. It remains unclear if the wide range of reliability estimates reported for MSSA are due to the qualitative nature of assessment, issues relating to study quality, or the wide range of study methodologies that include study population, patient position or examiner contact<sup>14,39,41,42,44</sup>. Systematic reviews uniformly reported that intra-rater reliability was consistently more favourable than inter-rater reliability<sup>39,41,43</sup>. The umbrella review (Chapter 2) highlighted that MSSA performed on a symptomatic population improved intra/inter-rater reliability<sup>41,43,44</sup>. This improvement in intra/inter-rater reliability is likely to be attributed to the patients' ability to recall and anticipate pain<sup>43,44</sup>. Conversely, disagreement surrounding segmental level was shown to decrease inter-rater reliability of MSSA<sup>41,43</sup>. Clinicians must be cognisant about factors that affect intra- and inter-rater reliability (limitations) of MSSA when applying techniques on patients in clinical practice.

In contrast to manual assessment, reviews that included mechanical and mechanically assisted devices used for the assessment of spinal stiffness demonstrated good test-retest reliability<sup>14,17</sup>. Unlike MSSA, there was no variation in test-retest reliability presented between reviews; clinically acceptable levels of test-retest reliability were reached indicating that mechanical and mechanically assisted devices used to assess spinal stiffness are more desirable for use in clinical or research environments, but are operationally more complex.

Chapter 2 supports the preferential use of mechanical and mechanically assisted where possible as reliability is quintessential for investigation of potential relationships between stiffness and other constructs (i.e. non-specific spinal pain).

### 5.1.2 Bench-top performance of the VerteTrack

The second aim of this thesis was to establish the bench-top performance of the VerteTrack, which was achieved via three discrete experiments to quantify the accuracy of (i) load applied by the device, (ii) displacement measured by the device, and (iii) stiffness of a test medium measured using dynamic and static indentation (Chapter 3 and 4).

Experiment one quantified the load applied by the VerteTrack ( $\text{Load}_{\text{VerteTrack}}$ ), which showed minimal measurement variance and high precision over repeated measures. The comparison between  $\text{Load}_{\text{VerteTrack}}$  and the same load measured externally ( $\text{Load}_{\text{ref}}$ ) was visualised using a Bland-Altman plot, which revealed a small negative systematic bias indicating that the load applied by the VerteTrack was consistently slightly less than the reference load (average bias of -0.05N, equivalent to 5 grams). Post-hoc analysis (after clear outliers were removed) demonstrated a reduction in the magnitude of bias by approximately 50%. It is unlikely that the detected difference would be clinically meaningful as it represents a systematic error of <0.1% (at 60N).

Experiment two quantified displacement measured by the VerteTrack ( $\text{Displacement}_{\text{VerteTrack}}$ ) over repeated measurements, which demonstrated minimal measurement variance and high precision. Displacement measured by the VerteTrack was compared to reference measurements, which demonstrated no statistically significant difference. However, the magnitude of the measurements were at the limit of resolution for both the string potentiometer in the VerteTrack (0.02mm) and the digital calliper used as a reference (0.05mm).

Similar to experiment one and two the coefficient of variation was low for dynamic indentation and static indentation. Both dynamic and static indentation demonstrated high repeatability and minimal measurement variance. There was small negative systematic bias indicating that the stiffness measured by dynamic indentation was marginally lower compared to static indentation (-0.25N/mm). Although the findings indicate a statistically

significant difference between dynamic and static indentation, it is unclear whether this difference is clinically significant. There is currently no published data to support a minimal clinically important difference (MCID) for the assessment of spinal stiffness<sup>14,17</sup>. More broadly mechanical devices need to collect baseline spinal stiffness data in a human population in order to determine a MCID. This will allow for more robust conclusions regarding differences between static and dynamic methods of indentation. Albeit, during a clinical trial to standardise procedure only one method of indentation should be used throughout so potential differences would not affect the results. Confirmation of these discrepancies between indentation methods should be sought on a human population.

The plot of raw dynamic and static stiffness values demonstrated that lower loads (RIH+1 and RIH+2) exhibit higher stiffness values, suggesting that lower loads may not provide valid estimates of stiffness. Increased stiffness values may be due to lower load not being able to sufficiently overcome the initial friction of the medium. This phenomenon is likely attributed to the medium rather than the device, as the inverse is seen in participants (where lower loads yield lower stiffness values). Indentation on participants demonstrates that lower loads frequently exhibit large quantities of initial displacement, is described as the ‘toe region’ in the literature<sup>17</sup>. Therefore, the utility of performing indentation at lower loads may be more time consuming while adding no additional information to the assessment of spinal stiffness. Commencing indentation at RIH+3 (approximately 50N) would reduce the assessment time by approximately half.

## **5.2 Strengths**

### **5.2.1 Reliability of existing spinal stiffness assessment methods**

In relation to the first aim of the thesis, Chapter 2 was the first umbrella systematic review that investigated the reliability of manual, mechanical and mechanically assisted test systems used to assess spinal stiffness. Reliability was extracted from nine reviews which identified 74 unique primary studies. This review provided: (i) a comprehensive summary of the available reviews, (ii) established agreement and disagreement among the literature, and (iii) exposed deficits in the current body of knowledge. The umbrella review also highlighted methodological limitations in the field and factors that affect reliability. The summation of included reviews provided commentary on the methodological challenges in the field, by highlighting heterogeneity which in part may contribute to the overall variable reliability of

MSSA<sup>14,17,18,39-44</sup>. Sources of primary study heterogeneity included the participant population (asymptomatic/symptomatic), participant position, practitioner contact and the reporting scale. The umbrella review informs clinicians and researchers about the utility of test systems employed to assess spinal stiffness.

### 5.2.2 Bench-top performance of the VerteTrack

In relation to the second aim, this is the first mechanical device to be assessed for bench-top performance prior to undertaking a test-retest reliability study. Inferences made regarding the accuracy of the device can be attributed to the device or medium, rather than due to the variability of a human population. This thesis quantifies differences between dynamic and static indentation. These findings in conjunction with previous research indicate that lower loads may not provide accurate stiffness estimates<sup>17</sup>. This thesis provides valuable information that may modify future protocols employed for the use of the VerteTrack (e.g. starting the load at RIH+3 on participants).

## **5.3 Limitations**

### 5.3.1 Reliability of existing spinal stiffness assessment methods

The limitations of the umbrella review performed in Chapter 2 were primarily due to the heterogeneity of reviews, particularly in relation to primary quality appraisal, data extraction and categorisation. First, a number of quality appraisal tools were used by included reviews, either as modified or combination of existing tools (e.g. QUADAS, QAERL, STARD)<sup>40,42,44</sup>, invention of new quality assessment tools<sup>18,39,43</sup>, or did not appraise primary study quality<sup>14,17,41</sup>. This limited the ability to compare recommendations between reviews. Second, reviews used different methods to extract data from the included studies and different statistics were reported by primary studies, which ultimately resulted in a range of statistical techniques such as; Cohens' kappa, intraclass correlation coefficient and percentage agreement. Last, the reviews used heterogeneous terminology to describe spinal stiffness (e.g. hypomobility, motion palpation and PAIVM) which meant that relevant reviews may have been missed from the search.

Further, there is no validated critical appraisal tool to assess the methodological quality for systematic reviews of reproducibility studies. In the absence of a specific tool, modification

to an existing validated assessment tool (AMSTAR-2, without a summary score) was judged to be the best option<sup>48</sup>. However, there were limitations in applying this tool.

### 5.3.2 Bench-top performance of the VerteTrack

Despite promising results with regard to bench-top accuracy, there are limitations that impact the generalisability of results from this study. This study was performed on a viscoelastic foam medium, without the presence of physiological properties known to influence spinal stiffness (such as breathing, spinal extensor muscle contraction and abdominal muscle contraction)<sup>14,17</sup>. In addition, it is unclear to what extent the observed phenomena can be attributed to the medium and whether a human population would emulate similar findings.

To quantify bias, the level of agreement between dynamic stiffness measurements were compared to a reference standard. Static indentation was used as a proxy reference standard given that is the more established method of indentation reported in the literature.

Unfortunately, there exists no ‘gold standard’ to ascertain spinal stiffness in human participants.

A high level of agreement was observed between both static and dynamic methods of stiffness assessment, despite a small negative systematic bias for dynamic compared to static indentation (-0.25N/mm). It is possible that a small systematic baseline offset error may be present in the LabView software (or measurement protocol) that reports Z-axis data for either method of indentation. It is also possible that dynamic (rolling) assessment of the test medium produces a different magnitude of stiffness compared to static assessment, due to the interface between the RIH and medium. In addition, the contact surface area of the device may vary according to the amount of force applied. For instance, higher loads would cause a larger surface area of the medium to be contacted by the rolling indenter head. These potential sources of error warrant further investigation in future studies.

## **5.4 Future directions**

Chapter 2 describes heterogeneity among primary reliability studies. Recommendations to improve future reliability studies include the recruitment of symptomatic populations and to achieve procedural standardisation (with regard to patient position, practitioner contact and

reporting). The umbrella review highlighted inconsistency in reliability for MSSA, bringing into question the veracity of MSSA as a clinical research tool. With regard to the reliability of mechanical devices, there is the potential for future research to explore the criterion validity of MSSA against mechanical methods of assessment. There are currently three studies that have compared MSSA against a mechanical reference standard to demonstrate poor correlation between the two methods<sup>22-24</sup>. Although current mechanical test systems have greater reliability compared to MSSA, there are feasibility issues in relation to high cost, increased assessment time and participant comfort to consider.

The VerteTrack offers a solution to reduce the cost and time of assessment, without compromising patient comfort<sup>35</sup>. This thesis investigated the bench-top performance of this device, which demonstrated both high precision and low systematic bias, highlighting that the VerteTrack has appropriate levels of accuracy to measure spinal stiffness. Moreover, the test-retest reliability of the VerteTrack must be determined in a human population. In the future, discrepancies between indentation methods may be confirmed by completing both methods of indentation on human subjects.

## **5.5 Concluding statement**

In conclusion, the findings of this thesis provide original contributions to the field of spinal stiffness assessment. An umbrella review highlighted the variable reliability of MSSA and demonstrated that mechanical and mechanically assisted devices have greater reliability. The umbrella review also highlighted feasibility issues in the implementation of mechanical devices into clinical practice and research settings. The Vertetrack demonstrated acceptable bench-top performance; exhibiting high precision, linearity and low systematic bias compared to reference values. The VerteTrack allows for dynamic indentation, which provides a comfortable solution to reduce assessment time. While discrepancies between dynamic and static spinal stiffness were found, estimates were small and unlikely to be clinically relevant. Albeit, this interpretation needs to be approached with caution; future research is now needed to assess the test-retest reliability of dynamic spinal stiffness assessment in human subjects.



## References

1. Ahmad Kiadaliri A, Wang H, Abajobir A, Abate K, Abbafati C, Abbas K, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*. 2017;390(10100):1260-344.
2. Hurwitz EL, Randhawa K, Yu H, Côté P, Haldeman S. The Global Spine Care Initiative: a summary of the global burden of low back and neck pain studies. *European Spine Journal*. 2018.
3. Walker B, Muller R, Grant W. Low back pain in Australian adults: the economic burden. *Asia Pacific Journal of Public Health*. 2003;15(2):79-87.
4. Refshauge KM, Maher CG. Low back pain investigations and prognosis: a review. *British Journal of Sports Medicine*. 2006;40(6):494-8.
5. Croft PR, Papageorgiou AC, Ferry S, Thomas E, Jayson M, Silman AJ. Psychologic distress and low back pain. Evidence from a prospective study in the general population. *BMJ*. 1995;20(24):2731-7.
6. Clark S, Horton R. Low back pain: a major global challenge. *The Lancet*. 2018;391(10137):2302.
7. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *The Lancet*. 2017;389(10070):736-47.
8. Cramer GD, Ross K, Raju P, Cambron J, Cantu JA, Bora P, et al. Quantification of cavitation and gapping of lumbar zygapophyseal joints during spinal manipulative therapy. *Journal of Manipulative & Physiological Therapeutics*. 2012;35(8):614-21.
9. Pickar JG. Neurophysiological effects of spinal manipulation. *The Spine Journal*. 2002;2(5):357-71.
10. Bialosky JE, Bishop MD, Price DD, Robinson ME, George SZ. The mechanisms of manual therapy in the treatment of musculoskeletal pain: A comprehensive model. *Manual Therapy*. 2009;14(5):531-8.
11. Qaseem A, Wilt TJ, McLean RM, Forciea MA. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. *Annals of Internal Medicine*. 2017;166(7):514-30.
12. Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: A validation study. *Annals of Internal Medicine*. 2004;141(12):920-8.
13. Maitland GD. *Vertebral manipulation*: Butterworth-Heinemann; 2013.
14. Snodgrass SJ, Haskins R, Rivett DA. A structured review of spinal stiffness as a kinesiological outcome of manipulation: its measurement and utility in diagnosis, prognosis and treatment decision-making. *Journal of Electromyography and Kinesiology*. 2012;22(5):708-23.
15. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*. 2011;48(6):661-71.
16. Abbott JH, Flynn TW, Fritz JM, Hing WA, Reid D, Whitman JM. Manual physical assessment of spinal segmental motion: Intent and validity. *Manual Therapy*. 2007;14(1):36.
17. Wong AYL, Kawchuk GN. The Clinical Value of Assessing Lumbar Posteroanterior Segmental Stiffness: A Narrative Review of Manual and Instrumented Methods. *PM & R*. 2017;9(8):816-30.

18. Seffinger MA, Najm WI, Mishra SI, Adams A, Dickerson VM, Murphy LS, et al. Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. *Spine*. 2004;29(19):413-25.
19. Nicholson LL, Maher CG, Adams R. Hand contact area, force applied and early non-linear stiffness (toe) in a manual stiffness discrimination task. *Manual Therapy*. 1998;3(4):212-9.
20. Maher C, Adams R. A Comparison of Pisiform and Thumb Grips in Stiffness Assessment. *Physical Therapy*. 1996;76(1):41-8.
21. Lee M, Svensson NL. Measurement of stiffness during simulated spinal physiotherapy. *Clinical Physics and Physiological Measurement*. 1990;11(3):201.
22. Maher CG, Latimer J, Adams R. An investigation of the reliability and validity of posteroanterior spinal stiffness judgments made using a reference-based protocol. *Physical Therapy*. 1998;78(8):829-37.
23. Koppenhaver SL, Hebert JJ, Kawchuk GN, Childs JD, Teyhen DS, Croy T, et al. Criterion validity of manual assessment of spinal stiffness. *Manual Therapy*. 2014;19(6):589-94.
24. Chiradejnant A, Maher CG, Latimer J. Objective manual assessment of lumbar posteroanterior stiffness is now possible. *Journal of Manipulative & Physiological Therapeutics*. 2003;26(1):34-9.
25. Stanton TR, Kawchuk GN. Reliability of assisted indentation in measuring lumbar spinal stiffness. *Manual Therapy*. 2009;14(2):197-205.
26. Kumar S, Stoll S. Device, protocol and measurement of regional spinal stiffness. *Journal of Electromyography and Kinesiology*. 2011;21(3):458-65.
27. Lee R, Evans J. Load-displacement-time characteristics of the spine under posteroanterior mobilisation. *Australian Journal of Physiotherapy*. 1992;38(2):115-23.
28. Latimer J, Goodsell MM, Lee M, Maher CG, Wilkinson BN, Moran CC. Evaluation of a new device for measuring responses to posteroanterior forces in a patient population, Part 1: Reliability testing. *Physical Therapy*. 1996;76(2):158-65.
29. Edmondston S, Allison G, Gregg C, Purden S, Svansson G, Watson A. Effect of position on the posteroanterior stiffness of the lumbar spine. *Manual Therapy*. 1998;3(1):21-6.
30. Brodeur RR, DelRe L. Stiffness of the thoracolumbar spine for subjects with and without low back pain. *Journal of the Neuromusculoskeletal System*. 1999;7(4):127-33.
31. Wong AY, Kawchuk G, Parent E, Prasad N. Within-and between-day reliability of spinal stiffness measurements obtained using a computer controlled mechanical indenter in individuals with and without low back pain. *Manual Therapy*. 2013;18(5):395-402.
32. Owens EF, Jr., DeVocht JW, Wilder DG, Gudavalli MR, Meeker WC. The Reliability of a Posterior-to-Anterior Spinal Stiffness Measuring System in a Population of Patients With Low Back Pain. *Journal of Manipulative & Physiological Therapeutics*. 2007;30(2):116-23.
33. Björnsdóttir SV, Guðmundsson G, Auðunsson GA, Matthíasson J, Ragnarsdóttir M. Posterior-anterior (PA) pressure Puffin for measuring and treating spinal stiffness: Mechanism and repeatability. *Manual Therapy*. 2016;22:72-9.
34. Kawchuk GN, Fauvel OR. Sources of variation in spinal indentation testing: Indentation site relocation, intraabdominal pressure, subject movement, muscular response, and stiffness estimation. *Journal of Manipulative & Physiological Therapeutics*. 2001;24(2):84-91.
35. Brown BT, Blacke A, Carroll V, Graham PL, Kawchuk G, Downie A, et al. The comfort and safety of a novel rolling mechanical indentation device for the measurement of lumbar trunk stiffness in young adults. *Chiropractic and Manual Therapies*. 2017;25(1).

36. ISO. 5725-1. Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions. 1 ed: International Organization for Standardization Geneva, Switzerland; 1998.
37. NATA. General Accreditation Guidance — Validation and verification of quantitative and qualitative test methods. 2018.
38. Menditto A, Patriarca M, Magnusson B. Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance*. 2007;12(1):45-7.
39. Haneline MT, Cooperstein R, Young M, Birkeland K. Spinal motion palpation: a comparison of studies that assessed intersegmental end feel vs excursion. *Journal of Manipulative & Physiological Therapeutics*. 2008;31(8):616-26.
40. Hollerwöger D. Methodological quality and outcomes of studies addressing manual cervical spine examinations: A review. *Manual Therapy*. 2006;11(2):93-8.
41. Huijbregts PA. Spinal motion palpation: A review of reliability studies. *Journal of Manual & Manipulative Therapy*. 2002;10(1):24.
42. Jonsson A, Rasmussen-Barr E. Intra- and inter-rater reliability of movement and palpation tests in patients with neck pain: A systematic review. *Physiotherapy Theory and Practice*. 2018;34(3):165-80.
43. Stockkendahl MJ, Christensen HW, Hartvigsen J, Vach W, Haas M, Hestbaek L, et al. Manual examination of the spine: a systematic critical literature review of reproducibility. *Journal of Manipulative & Physiological Therapeutics*. 2006;29(6):475-85, 85.e1-10.
44. van Trijffel E, Anderegg Q, Bossuyt PM, Lucas C. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Manual Therapy*. 2005;10(4):256-69.
45. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *International Journal of Evidence-Based Healthcare*. 2015;13(3):132-40.
46. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009;339:b2700.
47. 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.
48. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:4008.
49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;159-74.
50. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation*. 1998;12(3):187-99.
51. Johnston W. Inter-examiner reliability in palpation. *Journal of the American Osteopathic Association*. 1976;76:286-7.
52. Fritz JM, Koppenhaver SL, Kawchuk GN, Teyhen DS, Hebert JJ, Childs JD. Preliminary investigation of the mechanisms underlying the effects of manipulation: Exploration of a multivariate model including spinal stiffness, multifidus recruitment, and clinical findings. *Spine*. 2011;36(21):1772-81.
53. Hu Y, Wong YL, Lu WW, Kawchuk GN. Creation of an asymmetrical gradient of back muscle activity and spinal stiffness during asymmetrical hip extension. *Clinical Biomechanics*. 2009;24(10):799-806.
54. Kawchuk GN, Edgecombe TL, Wong AYL, Cojocaru A, Prasad N. A non-randomized clinical trial to assess the impact of nonrigid, inelastic corsets on spine function

in low back pain participants and asymptomatic controls. *The Spine Journal*. 2015;15(10):2222-7.

55. Tuttle N, Barrett R, Laakso L. Postero-anterior movements of the cervical spine: repeatability of force displacement curves. *Manual Therapy*. 2008;13(4):341-8.
56. Kumar S. Posteroanterior Spinal Stiffness at T5, T10, and L3 Levels in Normal Subjects. *PM & R*. 2012;4(5):342-8.
57. Hanneman SK. Design, Analysis and Interpretation of Method-Comparison Studies. *AACN advanced critical care*. 2008;19(2):223-34.
58. St-Pierre N. Validating mathematical models of biological systems: application of the concordance correlation coefficient. *Sensitivity Analysis of Model Output Los Alamos National Laboratory*. 2005:341-50.
59. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report: HAM2005-062*. 2005.
60. Stanton TR, Moseley GL, Wong AY, Kawchuk GN. Feeling stiffness in the back: a protective perceptual inference in chronic back pain. *Scientific reports*. 2017;7(1):9681.
61. Bergstrom E. An inter and intra-examiner reliability study of motion palpation of the lumbar spine in lateral flexion in the seated position. *European Journal of Chiropractic*. 1985.
62. Binkley J, Stratford PW, Gill C. Interrater reliability of lumbar accessory motion mobility testing. *Physical Therapy*. 1995;75(9):786-92.
63. Brismée J-M, Gipson D, Ivie D, Lopez A, Moore M, Matthijs O, et al. Interrater reliability of a passive physiological intervertebral motion test in the mid-thoracic spine. *Journal of Manipulative & Physiological Therapeutics*. 2006;29(5):368-73.
64. Bronemo L VSJ. A comparison of the inter- and intraexaminer reliability of motion palpation of the lower cervical spine (C2-C7) in the oblique-posterior lateral direction in sitting and supine positions. Thesis. Bournemouth, UK:Anglo-European College of Chiropractic. 1987.
65. Carmichael JP. Inter-and intra-examiner reliability of palpation for sacroiliac joint dysfunction. *Journal of Manipulative & Physiological Therapeutics*. 1987;10(4):164-71.
66. Christensen HW, Vach W, Vach K, Manniche C, Haghfelt T, Hartvigsen L, et al. Palpation of the upper thoracic spine: An observer reliability study. *Journal of Manipulative & Physiological Therapeutics*. 2002;25(5):285-92.
67. Cleland JA, Childs JD, Fritz JM, Whitman JM. Interrater reliability of the history and physical examination in patients with mechanical neck pain. *Archives of physical medicine and rehabilitation*. 2006;87(10):1388-95.
68. Comeaux Z, Eland D, Chila A, Pheley A, Tate M. Measurement challenges in physical diagnosis: refining inter-rater palpation, perception and communication. *Journal of Bodywork & Movement Therapies*. 2001;5(4):245-53.
69. Deboer K, Harmon JR, Tuttle C, Wallace H. Reliability study of detection of somatic dysfunctions in the cervical spine. *Journal of Manipulative & Physiological Therapeutics*. 1985;8(1):9-16.
70. Degenhardt BF, Snider KT, Snider EJ, Johnson JC. Interobserver reliability of osteopathic palpatory diagnostic tests of the lumbar spine: improvements from consensus training. *The Journal of the American Osteopathic Association*. 2005;105(10):465-73.
71. Downey B, Taylor N, Niere K. Can manipulative physiotherapists agree on which lumbar level to treat based on palpation? *Physiotherapy*. 2003;89(2):74-81.
72. Fjellner A, Bexander C, Faleij R, Strenger L-E. Interexaminer reliability in physical examination of the cervical spine. *Journal of Manipulative & Physiological Therapeutics*. 1999;22(8):511-6.
73. Gonnella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. *Physical Therapy*. 1982;62(4):436-44.

74. Grant A. An inter-and intra-examiner reliability study, using lateral flexion motion palpation of the lumbar spine in the prone position: Anglo-European College of Chiropractic; 1985.
75. Haas M. Reliability of manual end-play palpation of the thoracic spine. *Chiro Tech*. 1995;7:120-4.
76. Hanney WJ, George SZ, Kolber MJ, Young I, Salamh PA, Cleland JA. Inter-rater reliability of select physical examination procedures in patients with neck pain. *Physiotherapy Theory & Practice*. 2014;30(5):345-52.
77. Hanten WP, Olson SL, Ludwig GM. Reliability of manual mobility testing of the upper cervical spine in subjects with cervicogenic headache. *Journal of Manual & Manipulative Therapy*. 2002;10(2):76-82.
78. Heiderscheit B, Boissonnault W. Reliability of joint mobility and pain assessment of the thoracic spine and rib cage in asymptomatic individuals. *Journal of Manual & Manipulative Therapy*. 2008;16(4):210-6.
79. Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Archives of physical medicine and rehabilitation*. 2003;84(12):1858-64.
80. Horneij E, Hemborg B, Johnsson B, Ekdahl C. Clinical tests on impairment level related to low back pain: a study of test reliability. *Journal of Rehabilitation Medicine*. 2002;34(4):176-82.
81. Lynette Inscoe E, Witt PL, Gross MT, Mitchell RU. Reliability in Evaluating Passive Intervertebral Motion of the Lumbar Spine. *Journal of Manual & Manipulative Therapy*. 1995;3(4):135-43.
82. Johnston W, Hill J, Sealey J, Sucher B. Palpatory findings in the cervicothoracic region: variations in normotensive and hypertensive subjects. A preliminary report. *The Journal of the American Osteopathic Association*. 1980;79(5):300.
83. Johnston W, Hill J, Elkiss M, Marino R. Identification of stable somatic findings in hypertensive subjects by trained examiners using palpatory examination. *The Journal of the American Osteopathic Association*. 1982;81(12):830-6.
84. Johnston W, Beal M, Blum G, Hendra J, Neff D, Rosen M. Passive gross motion testing: Part III. Examiner agreement on selected subjects. *The Journal of the American Osteopathic Association*. 1982;81(5):309-13.
85. Johnston W, Elkiss M, Marino R, Blum G. Passive gross motion testing: Part II. A study of interexaminer agreement. *The Journal of the American Osteopathic Association*. 1982;81(5):304.
86. Jull G, Bullock M. A motion profile of the lumbar spine in an ageing population assessed by manual examination. *Physiotherapy Practice*. 1987;3(2):70-81.
87. Jull G, Zito G, Trott P, Potter H, Shirley D, Richardson C. Inter-examiner reliability to detect painful upper cervical joint dysfunction. *Australian Journal of Physiotherapy*. 1997;43(2):125-9.
88. Keating JJ, Bergmann T, Jacobs G, Finer B, Larson K. Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. *Journal of Manipulative & Physiological Therapeutics*. 1990;13(8):463-70.
89. Larsson A-C. A Pilotstudy to Compare the Intra-and Interreliability of Examiners Using Gillet's Motion Palpation Methods in the Lumbar Spine: Anglo-European College of Chiropractic; 1984.
90. Leboeuf-Yde C, Van Dijk J, Franz C, Hustad SA, Olsen D, Pihl T, et al. Motion palpation findings and self-reported low back pain in a population-based study sample. *Journal of Manipulative & Physiological Therapeutics*. 2002;25(2):80-7.

91. Lindsay DM, Meeuwisse WH, Mooney ME, Summersides J. Interrater reliability of manual therapy assessment techniques. *Physiotherapy Canada*. 1995;47(3):173-80.
92. Loram A. A comparative study of fixation findings in the thoracic spine in the sagittal plane using motion palpation in the sitting position and joint springing in the prone position: Anglo-European College of Chiropractic; 1987.
93. Love RM, Brodeur R. Inter-and intra-examiner reliability of motion palpation for the thoracolumbar spine. *Journal of Manipulative & Physiological Therapeutics*. 1987;10(1):1-4.
94. Lundberg G, Gerdle B. The relationships between spinal sagittal configuration, joint mobility, general low back mobility and segmental mobility in female homecare personnel. *Scandinavian journal of rehabilitation medicine*. 1999;31(4):197-206.
95. Maher C, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Physical Therapy*. 1994;74(9):801-9.
96. Marcotte J NM. Standardizing dynamic palpation in chiropractic: a reliability study for treatment of the neck area [in French]. *The Journal of the Canadian Chiropractic Association*. 2001;45:106–12.
97. Marcotte J, Normand MC, Black P. The kinematics of motion palpation and its effect on the reliability for cervical spine rotation. *Journal of Manipulative & Physiological Therapeutics*. 2002;25(7):E1-E9.
98. Mastriani PJ. Reliability of passive lumbar segmental motion. Boston, MA: MGH Institute of Health Professions. 1991.
99. McPartland JM, Goodridge JP. Counterstrain and traditional osteopathic examination of the cervical spine compared. *Journal of Bodywork & Movement Therapies*. 1997;1(3):173-8.
100. Mior SA, King RS, McGregor M, Bernard M. Intra and interexaminer reliability of motion palpation in the cervical spine. *The Journal of the Canadian Chiropractic Association*. 1985;29(4):195.
101. Mootz R, Keating JJ, Kontz H, Milus T, Jacobs G. Intra-and interobserver reliability of passive motion palpation of the lumbar spine. *Journal of Manipulative & Physiological Therapeutics*. 1989;12(6):440-5.
102. Nansel DD, Peneff AL, Jansen R, Cooperstein R. Interexaminer concordance in detecting joint-play asymmetries in the cervical spines of otherwise asymptomatic subjects. *Journal of Manipulative & Physiological Therapeutics*. 1989;12(6):428-33.
103. Olson KA, Paris SV, Spohr C, Gorniak G. Radiographic assessment and reliability study of the craniovertebral sidebending test. *Journal of Manual & Manipulative Therapy*. 1998;6(2):87-96.
104. Phillips DR TL. Comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. *Integrating Approaches Proceedings of the Eighth Biennial Conference of the Manipulative Physiotherapists Association of Australia*. 1993;55-61.
105. Phillips DR, Twomey L. A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. *Manual Therapy*. 1996;1(2):82-7.
106. Piva SR, Erhard RE, Childs JD, Browder DA. Inter-tester reliability of passive intervertebral and active movements of the cervical spine. *Manual Therapy*. 2006;11(4):321-30.
107. Pool JJ, Hoving JL, de Vet HC, van Mameren H, Bouter LM. The interexaminer reproducibility of physical examination of the cervical spine. *Journal of Manipulative & Physiological Therapeutics*. 2004;27(2):84-90.
108. Rhudy T, Sandefur M, Burk J. Interexaminer/intertechnique reliability in spinal subluxation assessment: a multifactorial approach. *American Journal of Chiropractic Medicine*. 1988;1:111-4.

109. Richter T, Lawall J. Reliability of diagnostic findings in manual medicine [in German][: Zur Zuverlässigkeit manualdiagnostischer Befunde]. *Manuelle Medizin*. 1993;31:1-11.
110. Schneider GM, Jull G, Thomas K, Smith A, Emery C, Faris P, et al. Intrarater and interrater reliability of select clinical tests in patients referred for diagnostic facet joint blocks in the cervical spine. *Archives of Physical Medicine & Rehabilitation*. 2013;94(8):1628-34.
111. Schoensee SK, Jensen G, Nicholson G, Gossman M, Katholi C. The effect of mobilization on cervical headaches. *Journal of Orthopaedic & Sports Physical Therapy*. 1995;21(4):184-96.
112. Schoeps P, Pfingsten M, Siebert U. Reliability of manual medical examination techniques of the cervical spine. Study of quality assurance in manual diagnosis. *Zeitschrift für Orthopädie und ihre Grenzgebiete*. 2000;138(1):2-7.
113. Sebastian D, Chovvath R. Reliability of palpation assessment in non-neutral dysfunctions of the lumbar spine. *Orthopaedic Physical Therapy Practice*. 2004;16:23-6.
114. Smedmark V, Wallin M, Arvidsson I. Inter-examiner reliability in assessing passive intervertebral motion of the cervical spine. *Manual Therapy*. 2000;5(2):97-101.
115. Smith AR CP, Nyberg RE. Intratester/intertester reliability of segmental motion testing of cervicothoracic forward bending in a symptomatic population. *Paris SV, Ed IFOMT Proceedings*. 1992;June 1-5:194.
116. Streder L-E, Sjöblom A, Sundell K, Ludwig R, Taube A. Interexaminer reliability in physical examination of patients with low back pain. *Spine*. 1997;22(7):814-20.
117. Streder L-E, Lundin M, Nell K. Interexaminer reliability in physical examination of the neck. *Journal of Manipulative & Physiological Therapeutics*. 1997;20(8):516-20.
118. Van Suijlekom HA, De Vet HC, Van Den Berg SG, Weber WE. Interobserver reliability in physical examination of the cervical spine in patients with headache. *Headache: The Journal of Head and Face Pain*. 2000;40(7):581-6.
119. Zito G, Jull G, Story I. Clinical tests of musculoskeletal dysfunction in the diagnosis of cervicogenic headache. *Manual Therapy*. 2006;11(2):118-29.

## **Appendix A: Search string**

### **PubMed**

03/04/18

((((((((((("Spinal stiffness") OR "spinal motion palpation") OR "spinal palpation") OR "intervertebral motion") OR "intersegmental motion") OR "posteroanterior stiffness") OR "lumbar stiffness") OR "thoracic stiffness") OR "cervical stiffness"))

### **Embase**

03/04/18

("Spinal stiffness" OR "spinal motion palpation" OR "spinal palpation" OR "intervertebral motion" OR "intersegmental motion" OR "posteroanterior stiffness" OR "lumbar stiffness" OR "thoracic stiffness" OR "cervical stiffness").af.

### **Cinahl**

03/04/18

((((((((((("Spinal stiffness") OR "spinal motion palpation") OR "spinal palpation") OR "intervertebral motion") OR "intersegmental motion") OR "posteroanterior stiffness") OR "lumbar stiffness") OR "thoracic stiffness") OR "cervical stiffness"))



## Appendix B: AMSTAR-2 Domains

AMSTAR-2 Domains	
1	Did the research questions and inclusion criteria for the review include the components of PICO?
2	Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?
3	Did the review authors explain their selection of the study designs for inclusion in the review?
4	Did the review authors use a comprehensive literature search strategy?
5	Did the review authors perform study selection in duplicate?
6	Did the review authors perform data extraction in duplicate?
7	Did the review authors provide a list of excluded studies and justify the exclusions?
8	Did the review authors describe the included studies in adequate detail?
9	Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?
10	Did the review authors report on the sources of funding for the studies included in the review?
11	If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?
12	If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?
13	Did the review authors account for RoB in primary studies when interpreting/discussing the results of the review?
14	Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?
15	If they performed quantitative synthesis, did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?
16	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?

## Appendix C. Primary reliability studies

Primary study	Haneline et al. 2008	Hollerwoger 2006	Huijbregts 2013	Jonsson et al. 2018	Seffinger et al. 2004	Snodgrass et al. 2012	Stochkendahl et al. 2006	Van Trijffel et al. 2005	Wong et al. 2017	Population	Number of participants	Region	Method category	Number of examiners	Examiner discipline	Quality rating	Intra-examiner reliability	Inter-examiner reliability	Test-retest reliability	Reporting Method
Bergstroem and Courtis (1986) <sup>61</sup>	Y	Y			Y		Y	Y		-	100	L°	MSSA	2	CS	Haneline (intra 50% and inter 67%) Seffinger 55.5% Stochkendahl 0%	Haneline PA=91-100% Huijbregts PA=92.2-99.2% Seffinger PA=95.4%	Haneline PA=65-88% Huijbregts PA=65-88% Seffinger PA=81.8% Van Trijffel 73-89%	-	Dichotomous
Binkley et al. (1995) <sup>62</sup>	Y	Y			Y		Y			Sx	18	L°	MSSA	6	P	Haneline 50% Seffinger 47% Stochkendahl 33.3%	-	Haneline k=0.09 and ICC=0.25 (CI, 0-0.39) Huijbregts ICC 0.25 Seffinger ICC=0.25 Stochkendahl ICC=0.09-0.25	-	9-point scale
Bjornsdottir et al. (2016) <sup>33</sup>									Y	Asx	30	T° and L°	MA			-	-	-	Wong ICC=0.83	Quantitative
Boline et al. (1988)	Y	Y			Y		Y	Y		Sx/Asx	50	L°	MSSA	2	C/CS	Haneline 83% Seffinger 60% Stochkendahl 66.7%	-	Haneline k=-0.05-0.31 PA 60-90% Huijbregts k=0-0.27 PA 60-90% Seffinger k=-0.03-0.37 PA 90-96% Stochkendahl k=-0.05-0.31 PA 78-91% Van Trijffel k=-0.06-0.33 PA 60-90%	-	Dichotomous
Brismee et al. (2006) <sup>63</sup>	Y									Asx	41	T°	MSSA	3	P	Haneline 50%	-	Haneline k=0.27-0.65 PA 63-83%		Dichotomous
Brodeur and DelRe (1999) <sup>30</sup>					Y			Y		Sx/Asx	67	T° and L°	MA	-	-	-	-	-	Wong SEM= 4.5 +/- 2.1% Snodgrass -	Quantitative

Bronemo and Van Steveninck (1987) <sup>64</sup>	Y			-	102	C°	MSSA	3	CS	-	Huijbregts PA=88.2-94.7%	Huijbregts PA=84.4-84.8%	-	Dichotomous		
Carmichael et al. (1987) <sup>65</sup>	Y			Asx	54	L°	MSSA	10	S	Stochkendahl 50%	Stochkendahl k=0.31 PA 90%	Stochkendahl k=0.02 PA 85%	-	Dichotomous		
Christensen et al. (2002) <sup>66</sup>	Y	Y	Y	Asx	107	T°	MSSA	2	C	Haneline 100% Stochkendahl 100%	Haneline k=0.59-0.64 Stochkendahl k=0.13-0.45 PA=82-88%	Haneline k=0.22-0.24 Stochkendahl k=-0.03-0 PA 68-80% Van Trijfffel k=-0.42-0.11 PA 68-77%	-	Dichotomous		
Cleland et al. (2006) <sup>67</sup>	Y			Sx	22	C°	MSSA	4	P	Jonsson high RoB	-	Jonsson k=-0.26-0.74	-	Trichotomous		
Comeaux et al. (2001) <sup>68</sup>	Y	Y	Y	Asx	54	C° and T°	MSSA	3	O	Haneline 67% Seffinger 52.5% Stochkendahl 50%	-	Haneline k=0.12-0.56 Seffinger k=0.16-0.43	-	Dichotomous		
Deboer et al. (1985) <sup>69</sup>	Y	Y	Y	Asx	40	C°	MSSA	2	CS	Seffinger 64.5%	Huijbregts k(w)=0.01-0.76 PA=58-75% Seffinger k(w)=0.01-0.76 PA 45-75%	Huijbregts k(w)=0.01-0.45 PA=21-58% Seffinger k(w)=0.01-0.45 Van Trijfffel k(w)=-0.03-0.45 PA 21-58%	-	Trichotomous		
Degenhardt et al. (2005) <sup>70</sup>	Y			Asx	15	L°	MSSA	3	C	Haneline 50%	-	Haneline k=0.2 PA 66%	-	-		
Downey et al. (2003) <sup>71</sup>	Y		Y	Sx	Haneline (n=30) Stochkendahl (n=60)	L°	MSSA	6	P	Haneline 33% Stochkendahl 50%	-	Haneline k=0.23-0.54 Stochkendahl k=0.37	-	Dichotomous		
Edmondston et al. (1998) <sup>29</sup>	Y		Y	Asx	8	L°	M	-	-	-	-	Wong ICC=0.979	-	Quantitative		
Fjellner et al. (1999) <sup>72</sup>	Y	Y	Y	Y	Y	Sx/Asx	Haneline (n=48) Seffinger (n=47) Stochkendahl (n=48)	C° and T°	MSSA	2	P	Haneline 67% Seffinger 74% Stochkendahl 66.7%	-	Haneline k(w)=0.01-0.18 PA 60-87% Hollerwoger k=-0.15-0.05 PA 41-90% Seffinger k(w)>0.4 in 5/58 tests Stochkendahl k(w)=-0.16-0.49 PA 41-92% Van Trijfffel k(w)=-0.16-0.49 PA 62-92%	-	Trichotomous
Fritz et al. (2011) <sup>52</sup>			Y	Sx/Asx	244	L°	M	-	-	-	-	-	Wong ICC=0.98-0.99	Quantitative		

Gonella et al. (1982) <sup>73</sup>	Y		Y	Y		Asx	5	L°	MSSA	5	P	Stochkendahl (intra 0% and inter 16.7%)	Huijbregts "good to reasonable"	Huijbregts - Stochkendahl - Van Trijffel mean(SD)=2.05-3.23(0-0.76)	-	7 point scale
Grant and Spadon (1985) <sup>74</sup>	Y	Y				-	60	L°	MSSA	4	CS	Seffinger 65.5%	Huijbregts PA 85-90% Seffinger PA 85-90%	Huijbregts PA 52.5-84.7% Seffinger PA 66.7%	-	Dichotomous
Haas et al. (1995) <sup>75</sup>	Y	Y	Y	Y	Y	Sx/Asx	73	T°	MSSA	2	C	Haneline 67% Seffinger 64.5% Stochkendahl 100%	Huijbregts k= 0.43-0.55 Seffinger k=0.43-0.55	Haneline k=0.08-0.22 Huijbregts k= 0.14-0.35 Seffinger k=0.19 Stochkendahl k=0.14 Van Trijffel k=-0.04-0.03 (SE 0.01-0.03)	-	Dichotomous
Hanney et al. (2014) <sup>76</sup>	Y					Sx	22	C°	MSSA	6	P	Jonsson high RoB	-	Jonsson k=0.15-0.43	-	Trichotomous
Hanten et al. (2002) <sup>77</sup>	Y	Y				Sx	Haneline (inter n=40 and intra n=20) Jonsson (n=20)	C°	MSSA	2	P	Haneline (intra 25% and inter 50%) Jonsson high RoB	Haneline k=0.21-0.80 PA 60-90% Jonsson k=-0.05-1	Haneline k=0.71-0.86 PA 70-95% Jonsson k=0.21-1.0	-	Dichotomous
Heiderscheit and Boissonnault (2008) <sup>78</sup>		Y				Asx	9	T°	MSSA	2	P	-	Snodgrass 0.61-0.75	Snodgrass k=0.59	-	Trichotomous
Hicks et al. (2003) <sup>79</sup>	Y		Y	Y		Sx	63	L°	MSSA	4	P/ C	Haneline 33% Stochkendahl 50%	-	Haneline k=-0.02-0.26 PA 52-69% Stochkendahl k=-0.02-0.26 PA 52-69% Van Trijffel k=-0.02-0.26 PA 52-69%	-	Trichotomous
Horneij et al. (2002) <sup>80</sup>		Y				Sx/Asx	84	T° and L°	MSSA	3	P	Stochkendahl 66.7%	Stochkendahl k=0.56-0.78 PA 78-89%	Stochkendahl k=0.12-0.49 PA 61-77%	-	Hypermobility with P
Hu et al. (2009) <sup>53</sup>			Y			Asx	83	L°	MA	-	-	-	-	-	Wong ICC=0.91-0.93	Quantitative
Inscoc et al. (1995) <sup>81</sup>	Y	Y	Y	Y	Y	Sx	6	L°	MSSA	2	P	Seffinger 59% Stochkendahl (intra 0% and inter 16.7%)	Huijbregts PA 66.67-75% Seffinger PA 66.67% and 75% Snodgrass PA 67% and 75%	Huijbregts PA 48.61% Seffinger PA 48.61% Snodgrass PA 49% Van Trijffel PA 33.33-58.33%	-	Trichotomous

Johnston et al. (1976) <sup>51</sup>	Y	-	10	C° and T°	MSSA	3	O	Seffinger 30%	-	Seffinger PA 40-60%	-	-				
Johnston et al. (1980) <sup>82</sup>	Y	Asx	132	C° and T°	MSSA	3	O	Seffinger 67%	-	Seffinger PA 39.5%	-	-				
Johnston et al. (1982) a <sup>83</sup>	Y	-	307	C° and T°	MSSA	3	O	Seffinger 71%	-	Seffinger X^2= 27.75, df= 1, p < 0.001	-	-				
Johnston et al. (1982) b <sup>84</sup>	Y	-	70	C°	MSSA	5	O	Seffinger 56.5%	-	Seffinger sum of the mean=13.2-15.6, SD 2.0 and 3.5 and p<0.35	-	-				
Johnston et al. (1982) c <sup>85</sup>	Y	-	161	C°	MSSA	3	O	Seffinger 54%	-	Seffinger observed agreement=12-18, z=2.5-3.64, alpha=.0005-.03	-	-				
Jull and Bullock (1987) <sup>86</sup>	Y	Y	Y	Asx	Haneline (intra n=20 and inter n=10) Huijbregts (n=25) Snodgrass (n=200 pooled from 2 Jull and Bullock studies)	L°	MSSA	2	P	Haneline 0% Haneline r=0.81-0.98 PA 87.5% Huijbregts r=0.81-0.91 PA 87.5%	Haneline r=0.82-0.94 PA 86% Huijbregts r=0.82-0.94 PA 86%	-	5 point scale			
Jull et al. (1997) <sup>87</sup>	Y	Sx/Asx	40	C°	MSSA	7	P	-	-	Huijbregts k=0.25-1	-	Dichotomous				
Keating et al. (1990) <sup>88</sup>	Y	Y	Y	Y	Y	Sx/Asx	46	L°	MSSA	3	C	Haneline 67% Seffinger 67.5% Stochkendahl 75%	-	Haneline k=-0.18-0.31 Huijbregts k(m)=0.03 to 0.23 Seffinger k=-0.03-0.23 Stochkendahl k=0.07-0.09 Van Trijffel k=-0.18-0.31	-	Dichotomous
Larsson (1984) <sup>89</sup>	Y	Asx	32	L°	MSSA	4	C/CS	-	Huijbregts PA 66.6-78.6%	Huijbregts PA 56.4%	-	Dichotomous				

Latimer et al. (1996) <sup>28</sup>			Y		Y	Sx	22	L°	M	-	-	-	-	-	Snodgrass "reliable with repeated testing 5 minues apart" Wong ICC=0.95-0.99	Quantitative	
Lebouf et al. (1989) <sup>90</sup>	Y		Y			Sx	45	L°	MSSA	4	CS	Stochkendahl (intra 25% and inter 16.7%)	Huijbreegts PA 50-90%	Huijbregts 20-100%	-	-	
Lee and Evans (1992) <sup>27</sup>					Y	Asx	10	L°	M	-	-	-	-	-	Wong ICC=0.95-0.99	Quantitative	
Lee and Stevensen (1990) <sup>21</sup>			Y		Y	Asx	11	L°	M	-	-	-	-	-	Snodgrass ICC=0.99 Wong ICC=0.88	Quantitative	
Lindsay et al. (1995) <sup>91</sup>	Y		Y		Y	Sx/Asx	18 Seffinger (n=8) Stochkendahl (Asx n=8)	L°	MSSA	2	P	Haneline 100% Seffinger 35% Stoochkendahl 166.7%	-	Haneline k(w)=-0.03-0.6 PA 14-100% Seffinger k=-0.5-0.3 PA > 70% for 8/20 tests Stoochkendahl k=-0.3-0 PA 14-50%	-	Dichotomous	
Loram (1987) <sup>92</sup>	Y					-	10	T°	MSSA	1	CS	-	Huijbregts mean agreement 86.1-100%	-	-	Dichotomous	
Love and Brodeur (1987) <sup>93</sup>	Y	Y		Y	Y	Y	Asx	32	T° and L°	MSSA	8	CS	Haneline (intra 0% and inter 17%) Seffinger 72% Stochkendahl (intra 0% and inter 16.7%)	Haneline r=0.02-0.65 Huijbregts r=0.302-0.6856 Seffinger r=0.302-0.684 Snodgrass "consistent"	Haneline r=0.01-0.49 Huijbregts r=0.023-0.085 Seffinger r=0.023-0.0852 Snodgrass "poor" pearson's r>0.3	-	Dichotomous
Lundberg and Gerdle (1999) <sup>94</sup>			Y		Y		Sx/Asx	Seffinger (n=150) Stochkendahl (n=156)	T° and L°	MSSA	3	P	Seffinger 68% Stochkendahl 66.7%	-	Seffinger k(w)=0.42-0.75 Stochkendahl k(w)=0.42-0.75	-	5 point scale

Maher and Adams (1994) <sup>95</sup>	Y	Y	Y	Y	Y	Y	Sx	90	L°	MSSA	6	P	Haneline 67% Seffinger 66% Stochkendahl 58.3%	-	Haneline ICC=0.04-0.73 PA 13-43% Huijbregts ICC=0.03-0.37 Seffinger ICC=-0.04-0.73 PA 13-43% Snodgrass ICC=0.03-0.37 PA 21-29% Stochkendahl ICC=-0.4-0.73 Van Trijffel ICC=-0.40-0.73 PA 13-43%	-	11-point scale
Maher et al. (1998) <sup>22</sup>	Y	Y	Y	Y			Asx	40	L°	MSSA	2	P	Haneline 33% Seffinger 51.5%	-	Haneline ICC=0.5-0.77 Huijbregts ICC=0.5-0.77 Seffinger ICC=0.5-0.77 Snodgrass ICC=0.5-0.77	-	11-point scale
Marcotte et al. (2001) <sup>96</sup>			Y				Sx	12	C°	MSSA	3	C	Seffinger 58%	Seffinger k=0.78 PA 90.6%	Seffinger k=0.57-0.85 PA 82.3-93.2%	-	-
Marcotte et al. (2002) <sup>97</sup>	Y	Y			Y		Asx	3	C°	MSSA	25	C/CS	Haneline 30% Stochkendahl 16.7%	-	Haneline k=0.6-0.8 Hollerwoger k=0.337-0.682 PA 81-90% Stochkendahl k=0.337-0.682 PA 81-90%	-	Dichotomous
Mastriani et al. (1991) <sup>98</sup>			Y				Sx	16	L°	MSSA	3	P	Seffinger 61.5%	-	Seffinger PA 62-66%	-	-
McPartland and Goodridge (1997) <sup>99</sup>	Y				Y		Sx/Asx	18	C°	MSSA	2	O	Haneline 83% Stochkendahl 58.3%	-	Haneline k=0.34 PA 66.7% Stochkendahl k=0.34 PA 67%	-	10-point scale
Mior and King et al. (1985) <sup>100</sup>	Y	Y	Y		Y	Y	Asx	Haneline (n=59) Seffinger (n=59) Stochkendahl (n=62)	C°	MSSA	2	CS	Haneline 50% Seffinger 55.5% Stochkendahl 50%	Haneline k=0.37-0.52 PA 71-79% Huijbregts k=0.37-0.52 PA 71-79% Seffinger k=0.37-0.52 PA 71-79% Stochkendahl k=0.37-0.52 PA 71-79%	Haneline k=0.15 PA 61% Huijbregts k=0.15 PA 61% Seffinger k=0.15 PA 61% Stochkendahl k=0.15 PA 61% Van Trijffel k=0.15 PA 62%	-	Dichotomous

Mootz et al. (1989) <sup>101</sup>	Y	Y	Y	Y	Y	Sx/Asx	60	L°	MSSA	2	C	Haneline (intra 25% and inter 33%) Seffinger 55% Stochkendahl (intra 25% inter 33.3%)	Haneline k=-0.09-0.48 Huijbregts k=-0.09- 0.48 Seffinger k=-0.11-0.48 Stochkendahl k=-0.09- 0.48	Haneline k=-0.17-0.17 Huijbregts k=-0.17- 0.17 Seffinger k=-0.19-0.17 Stochkendahl k=-0.17- 0.17 Van Trijffel k=-0.17- 0.17 PA 61.7-85%	-	Dichotomous	
Nansel et al. (1989) <sup>102</sup>	Y	Y	Y	Y	Y	Asx	270	C°	MSSA	4	C/ CS	Haneline 50% Seffinger 58.5% Stochkendahl 16.7%	-	Haneline k=0.01 PA 45.6-54.3% Hollerwoger k=0.013 Huijbregts k=0.013 PA 50% Seffinger k=0.013 Stochkendahl k=0.01 PA 45.6-54.3%	-	Dichotomous	
Olson et al. (1998) <sup>103</sup>	Y			Y		Asx	10	C°	MSSA	6	P	Haneline (intra 25% and inter 33%) Seffinger 37.5%	Haneline k=0.01-0.31 Seffinger k=-0.22- 0.308	Haneline k=-0.04-0.12 Seffinger k=-0.043- 0.194	-	5-point scale (P) Quantitative (x- ray)	
Owens et al. (2007) <sup>32</sup>				Y		Y	Sx	L°	MA	-	-	-	-	-	Snodgrass ICC=0.79 Wong ICC=0.79	Quantitative	
Phillips and Twomey (1993) <sup>104</sup>		Y				Sx	72	L°	MSSA	2	P	-	-	Huijbregts k(w)=- 0.16-0.87 PA 30-100%	-	Dichotomous	
Phillips and Twomey (1996) <sup>105</sup>	Y	Y	Y			Sx/Asx	72	L°	MSSA	2	P	Haneline 67% Seffinger 63%	-	Haneline k(w)=-0.15- 0.24 PA 74-100% Huijbregts k(w)=- 0.14-0.24 PA 74-99% Seffinger k(w)=-0.15- 0.32 PA 55-100%	-	Trichotomous	
Piva et al. (2006) <sup>106</sup>		Y				Sx	30	C°	MSSA	2	P	Jonsson high RoB	-	Jonsson k=-0.07-0.81 PA 66-92%	-	Dichotomous	
Pool et al. (2004) <sup>107</sup>		Y	Y		Y	Y	Sx	32	C°	MSSA	2	P	Jonsson moderate RoB Stochkendahl 50%	-	Hollerwoger k=-0.1- 0.65 Jonsson k=-0.09-0.63 PA 68-84% Stochkendahl k=-0.09- 0.63 PA 48-90% Van Trijffel k=-0.09- 0.63 PA 68-90%	-	Dichotomous



Rhudy et al. (1988) <sup>108</sup>	Y										Sx	14	C°, T° and L°	MSSA	3	C	Seffinger 34%	-	Seffinger strength of agreement [(K score/sample size) X 100]: low=35%, substantial=11%, moderate=12%, medium=9%, almost perfect=8%, not observed=25%	-	-
Richter and Lawall (1993) <sup>109</sup>	Y	Y	Y	Y	Y	Y	Y	Y	Sx/Asx	61	L°	MSSA	5	M D	Seffinger 40%	Huijbregts k=0.3-0.8 Seffinger k=0.3-0.8	Huijbregts k=0.08-0.18 Seffinger k=0.08-0.47 Van Trijffel k=0.08-0.72	-	Trichotomous		
Scheider et al. (2013) <sup>110</sup>	Y								Sx	56	C°	MSSA	2	P	Jonsson low RoB	Jonsson k=0.62-0.88 PA 95%	Jonsson k=0.79-0.88 PA 98%	-	Dichotomous		
Schoensee et al. (1995) <sup>111</sup>	Y								Asx	10	C°	MSSA	2	P	-	Huijbregts k=0.81	Huijbregts k=0.45-0.79	-	Trichotomous		
Schoeps et al. (2000) <sup>112</sup>	Y	Y	Y	Y	Y	Y	Y	Y	Sx	20	C°	MSSA	5	M D	Seffinger 77.5%	-	Hollerwoger k=0.025-0.45 Seffinger k=0.2-0.4 Van Trijffel k=0.03-0.44	-	Dichotomous		
Sebastian and Chovvath (2004) <sup>113</sup>	Y								Sx	31	L°	MSSA	2	P	Stochkendahl 16.7%	-	Stochkendahl k=0.69	-	Dichotomous		
Smedmark and Wallin (2000) <sup>114</sup>	Y	Y	Y	Y	Y	Y	Y	Y	Sx	61	C°	MSSA	2	P	Haneline 67% Jonsson high RoB Seffinger 42% Stochkendahl 66.7%	-	Haneline k=0.28-0.43 PA 70-87% Hollerwoger k=0.28-0.45 PA 70-90% Huijbregts k=0.28-0.43 PA 70-87% Jonsson k=0.28-0.43 PA 70-87% Seffinger k=0.28-0.43 PA 70-87% Snodgrass k=0.28-0.43 PA 77% Stochkendahl k=0.28-0.43 PA 79-87% Van Trijffel k=0.28-0.43 PA 70-87%	-	Dichotomous		
Smith et al. (1992) <sup>115</sup>	Y								Sx	27	C° and T°	MSSA	3	P	-	Huijbregts k=0.291-1.00	Huijbregts k=0.057-0.602	-	7 point scale		

Stanton and Kawchuk (2009) <sup>25</sup>	Y	Y	Asx	23	L°	MA	-	-	-	-	-	Snodgrass ICC 0.91–0.93 Wong ICC 0.91–0.93	Quantitative		
Strender et al. (1997) a <sup>116</sup>	Y	Y	Y	Sx	50	L°	MSSA	4	M D/ P	Seffinger 62.5% Stochkendahl 66.7%	-	Seffinger k=-0.08-0.24 PA 48-86% Stochkendahl k=0.38-0.75 PA 72-88% Van Trijffel k(w)=0.66-0.75 PA 80-82%	-	Trichotomous	
Strender et al. (1997) b <sup>117</sup>	Y	Y	Y	Y	Sx/Asx	50	C°	MSSA	2	P	Seffinger 79% Stochkendahl 75%	-	Hollerwoger k=0.05-0.25 PA 44-71% Seffinger k=0.09-0.15 PA 26-44% Stochkendahl k=0.05-0.15 PA 26-44% Van Trijffel k=0.57-0.15 PA 26-44%	-	Dichotomous
Tuttle et al. (2008) <sup>55</sup>	Y			Asx	10	C°	M	-	-	-	-	-	Snodgrass coefficient of multiple determination 0.73-0.90	Quantitative	
Van Suijlekom et al. (2000) <sup>118</sup>	Y			Sx	24	C°	MSSA	2	M D	Seffinger 33.5%	-	Seffinger k=0.27-0.46	-	-	
Wong et al. (2013) <sup>31</sup>		Y		Sx/Asx	244	L°	M	-	-	-	-	-	Wong ICC=0.98-0.99	Quantitative	
Zito et al. (2006) <sup>119</sup>	Y			Sx	77	C°	MSSA	1	P	Haneline 75%	Haneline k=0.78-1 PA 70-87%	-	-	-	

Symptomatic (Sx), asymptomatic (Asx), symptomatic and asymptomatic (Sx/Asx), Assessment region: cervical (C°), thoracic (T°) and lumbar (L°), physiotherapist (P), chiropractor (C), osteopath (O), medical doctor (MD) and students (S) added to the end of the discipline when specified Assessment type: manual spinal stiffness assessment (MSSA), mechanical (M) and mechanically assisted (MA)

